Europäisches Patentamt | European Patent Office | Office européen des brevets

Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. | Patent application No. | Demande de brevet n°

02102626.5

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 02102626.5
Demande no:

Anmeldetag:
Date of filing: 22.11.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Koninklijke Philips Electronics N.V.
Groenewoudseweg 1
5621 BA Eindhoven
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Spracherkennungseinrichtung mit Mitteln zum Berücksichtigen von mindestens zwei Spracheigenschaften

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/ Classification internationale des brevets:

G10L15/26

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR

Spracherkennungseinrichtung mit Mitteln zum Berücksichtigen von mindestens zwei Spracheigenschaften

Die Erfindung bezieht sich auf eine Spracherkennungseinrichtung zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation.

Die Erfindung bezieht sich weiters auf ein Spracherkennungsverfahren zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation.

Die Erfindung bezeiht sich weiters auf ein Computerprogrammprodukt, das zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation ausgebildet ist.

Die Erfindung bezieht sich weiters auf einen Computer der das Computerprogrammprodukt gemäß dem vorstehenden Absatz abarbeitet.

Eine solche Spracherkennungseinrichtung der eingangs im ersten Absatz angeführten Gattung und ein solches Spracherkennungsverfahren der eingangs im zweiten Absatz angeführten Gattung und ein solches Computerprogrammprodukt der eingangs im dritten Absatz angeführten Gattung und ein solcher Computer der eingangs im vierten Absatz angeführten Gattung sind aus dem Patentdokument WO 98/08215 bekannt.

Bei der bekannten Spracherkennungseinrichtung sind Sprach-Erkennungsmittel vorgesehen, denen über ein Mikrofon eine Sprachinformation zugeführt wird. Die Spracherkennungsmittel sind unter fortwährender Berücksichtigung einer Eigenschaftsinformation, welche den jeweils beim Erkennen der Textinformation zu verwenden Kontext repräsentiert, zum Erkennen der Textinformation in der Sprachinformation ausgebildet. Zum Zweck des Erzeugens der Eigenschaftsinformation weist die Spracherkennungseinrichtung Spracheigenschaft-Erkennungsmittel auf, die zum Empfangen einer Repräsentation der Sprachinformation von den Sprach-Erkennungsmitteln und unter Ausnutzung der Repräsentation der Sprachinformation zum Erkennen des jeweils vorliegenden Kontexts als eine die Sprachinformation charakterisierende Spracheigenschaft und zum Erzeugen der den vorliegenden Kontext repräsentierenden Eigenschaftsinformation ausgebildet ist.

Bei der bekannten Spracherkennungseinrichtung besteht das Problem, dass zwar das Erkennen einer einzigen die Sprachinformation charakterisierenden Spracheigenschaft, nämlich das Erkennen des jeweils vorliegenden Kontexts, vorgesehen ist, jedoch andere die Sprachinformation charakterisierende Spracheigenschaften, wie eine Sprachsegmentierung oder die jeweils verwendete Sprache oder die jeweils vorliegende Sprechergruppe, während des Erkennens der Textinformation unberücksichtigt bleiben. Daher müssen diese unberücksichtigten Spracheigenschaften vor einem Einsatz der bekannten Spracherkennungseinrichtung vorbekannt sein und – für den Fall, dass ihnen überhaupt Rechnung getragen werden kann - gegebenenfalls fix voreingestellt, also unveränderbar vorkonfiguriert sein, wodurch jedoch der Einsatz der bekannten Spracherkennungseinrichtung bei einem Anwendungsfall, bei dem sich diese unberücksichtigbaren Spracheigenschaften während des Betriebs – also während des Erkennens der Textinformation - verändern, nicht möglich ist.

Die Erfindung hat sich zur Aufgabe gestellt, das vorstehend angeführte Problem bei einer Spracherkennungseinrichtung der eingangs im ersten Absatz angeführten Gattung und bei einem Spracherkennungsverfahren der eingangs im zweiten Absatz angeführten Gattung und bei einem Computerprogrammprodukt der eingangs im dritten Absatz angeführten Gattung und bei einem Computer der eingangs im vierten Absatz angeführten Gattung zu beseitigen und eine verbesserte Spracherkennungseinrichtung und ein verbessertes Spracherkennungsverfahren und ein verbessertes Computerprogrammprodukt und einen verbesserten Computer zu schaffen.

Zur Lösung der vorstehend angeführten Aufgabe sind bei einer Spracherkennungseinrichtung gemäß der Erfindung erfindungsgemäße Merkmale vorgesehen, so dass eine Spracherkennungseinrichtung gemäß der Erfindung auf die nachfolgend angegebene Weise charakterisierbar ist, nämlich:

Spracherkennungseinrichtung zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation, welche Sprachinformation hinsichtlich von Spracheigenschaften charakterisierbar ist, wobei erste Spracheigenschaft-Erkennungsmittel vorgesehen sind, die unter Ausnutzung der Sprachinformation zum Erkennen einer ersten Spracheigenschaft und zum Erzeugen einer die erkannte erste Spracheigenschaft

repräsentierenden ersten Eigenschaftsinformation ausgebildet sind, und wobei zumindest zweite Spracheigenschaft-Erkennungsmittel vorgesehen sind, die unter Ausnutzung der Sprachinformation zum Erkennen einer zweiten Spracheigenschaft der Sprachinformation und zum Erzeugen einer die erkannte zweite Spracheigenschaft repräsentierenden zweiten Eigenschaftsinformation ausgebildet ist, und wobei Sprach-Erkennungsmittel vorgesehen sind, die unter fortwährender Berücksichtigung von zumindest der ersten Eigenschaftsinformation und der zweiten Eigenschaftsinformation zum Erkennen der zu der Sprachinformation korrespondierenden Textinformation ausgebildet sind.

Zur Lösung der vorstehend angeführten Aufgabe sind bei einem Spracherkennungsverfahren gemäß der Erfindung erfindungsgemäße Merkmale vorgesehen, so dass ein Spracherkennungsverfahren gemäß der Erfindung auf die nachfolgend angegebene Weise charakterisierbar ist, nämlich:

Spracherkennungsverfahren zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation, welche Sprachinformation hinsichtlich von Spracheigenschaften charakterisierbar ist, wobei unter Ausnutzung der Sprachinformation eine erste Spracheigenschaft erkannt wird und wobei eine die erkannte erste Spracheigenschaft repräsentierende erste Eigenschaftsinformation erzeugt wird und wobei unter Ausnutzung der Sprachinformation mindestens eine zweite Spracheigenschaft erkannt wird und wobei eine die erkannte zweite Spracheigenschaft repräsentierende zweite Eigenschaftsinformation erzeugt wird und wobei die zu der Sprachinformation korrespondierende Textinformation unter fortwährender Berücksichtigung von zumindest der ersten Eigenschaftsinformation und der zweiten Eigenschaftsinformationen erkannt wird.

Zur Lösung der vorstehend angeführten Aufgabe ist bei einem Computerprogrammprodukt gemäß der Erfindung vorgesehen, dass das Computerprogrammprodukt direkt in einen Speicher eines Computers geladen werden kann und Softwarecodeabschnitte umfasst, wobei mit dem Computer das Spracherkennungsverfahren gemäß der Erfindung abgearbeitet werden kann, wenn das Computerprogrammprodukt auf dem Computer abgearbeitet wird.

Zur Lösung der vorstehend angeführten Aufgabe ist bei einem Computer gemäß der Erfindung vorgesehen, dass der Computer eine Recheneinheit und einen internen Speicher aufweist, der das Computerprogrammprodukt gemäß dem vorstehend

angeführten Absatz abarbeitet.

Durch das Vorsehen der Maßnahmen gemäß der Erfindung ist der Vorteil erhalten, dass ein zuverlässiges Erkennen einer Textinformation in einer Sprachinformation selbst bei einer Vielzahl von sich während des Erkennens der Textinformation verändernden Spracheigenschaften sichergestellt ist. Dadurch ist weiters der Vorteil erhalten, dass die Genauigkeit des Erkennens deutlich verbessert ist, weil ein durch ein Nichtberücksichtigen eines Veränderns einer Spracheigenschaft verursachtes Fehlerkennen der Textinformation durch das Erzeugen und Berücksichtigen der mindestens zwei Eigenschaftsinformationen auf zuverlässige Weise vermeidbar ist, da ein Verändern einer der Spracheigenschaften unmittelbar durch eine dieser Spracheigenschaft zugeordnete Eigenschaftsinformation repräsentiert wird und daher während des Erkennens der Textinformation berücksichtigbar ist. Dadurch ist weiters der Vorteil erhalten, dass durch die Vielzahl der zur Verfügung stehenden Eigenschaftsinformationen eine wesentlich genauere Modellierung der Sprache zum Erkennen der Textinformation verwendbar ist, was einen positiven Beitrag zur Genauigkeit des Erkennens der Spracheigenschaften und folglich auch zum Erkennen der Textinformation und weiters auch noch zur Geschwindigkeit des Erkennens der Textinformation liefert. Dadurch ist weiters der Vorteil erhalten, dass ein Einsatz der erfindungsgemäßen Spracherkennungseinrichtung in einem an die Flexibilität des Erkennens der Textinformation höchste Ansprüche stellenden Einsatzgebiet, wie beispielsweise bei einem Konferenz-Transkriptionssystem zum automatischen Transkribieren von einer bei einer Konferenz auftretenden Sprachinformation, ermöglicht ist. Bei diesem Einsatzgebiet ist sogar ein annähernd echtzeitmäßiges Erkennen der Textinformation selbst bei einem Vorliegen einer Sprachinformation realisierbar, die von unterschiedlichsten Sprechern mit unterschiedlichen Sprachen erzeugt wurde.

Bei den erfindungsgemäßen Lösungen hat es sich weiters als vorteilhaft erwiesen, wenn zusätzlich die Merkmale gemäß dem Anspruch 2 bzw. dem Anspruch 7 vorgesehen sind. Dadurch ist der Vorteil erhalten, dass die Bandbreite eines Audiosignals, das zum Empfangen der Sprachinformation eingesetzt wird, wobei die Bandbreite des Audiosignals von dem jeweiligen Empfangskanal abhängig ist, bei dem Erkennen der Eigenschaftsinformationen und/oder bei dem Erkennen der Textinformation berücksichtigbar ist.

Bei den erfindungsgemäßen Lösungen hat es sich weiters als vorteilhaft erwiesen, wenn zusätzlich die Merkmale gemäß dem Anspruch 3 bzw. dem Anspruch 8 vorgesehen sind. Dadurch ist der Vorteil erhalten, dass ein Teil der Sprachinformation erst dann von den Sprach-Erkennungsmitteln verarbeitet wird, wenn für diesen Teil der Sprachinformation gültige Eigenschaftsinformationen vorliegen, also die Spracheigenschaften für diesen Teil bestimmt wurden, so dass ein unnötiges Verschwenden oder Belegen von einer zum Erkennen der Textinformation benötigten Rechenleistung bzw. von sogenannten Systemressourcen zuverlässig vermeidbar ist.

Bei den erfindungsgemäßen Lösungen hat es sich weiters als vorteilhaft erwiesen, wenn zusätzlich die Merkmale gemäß dem Anspruch 4 bzw. dem Anspruch 9 vorgesehen sind. Dadurch ist der Vorteil erhalten, dass ein gegenseitiges Beeinflussen der mindestens zwei Spracheigenschaft-Erkennungsmittel ermöglicht ist. Dadurch ist weiters der Vorteil erhalten, dass ein sequentielles Erkennen der einzelnen Spracheigenschaften in einer das Erkennen der Spracheigenschaften begünstigenden Reihenfolge ermöglicht ist, was einen positiven Beitrag zu der Genauigkeit und der Geschwindigkeit des Erkennens der Textinformation leistet und eine verbesserte Ausnutzung von Rechenleistung ermöglicht.

Bei den erfindungsgemäßen Lösungen hat es sich weiters als vorteilhaft erwiesen, wenn zusätzlich die Merkmale gemäß dem Anspruch 5 bzw. dem Anspruch 10 vorgesehen sind. Dadurch ist der Vorteil erhalten, dass auf möglichst zuverlässige Weise ein Erkennen der jeweiligen Spracheigenschaft in Abhängigkeit von einer anderen Spracheigenschaft ermöglicht ist, weil die zum Erkennen der jeweiligen Spracheigenschaft ausnutzbare andere Spracheigenschaft erst dann verwendet wird, wenn die zu der anderen, also zu der zu berücksichtigenden Spracheigenschaft korrespondierende Eigenschaftsinformation tatsächlich verfügbar ist.

Bei einem erfindungsgemäßen Computerprogrammprodukt hat es sich weiters als vorteilhaft erwiesen, wenn zusätzlich die Merkmale gemäß dem Anspruch 11 vorgesehen sind. Dadurch ist der Vorteil erhalten, dass das Computerprogrammprodukt möglichst einfach vertrieben, verkauft oder vermietet werden kann.

Die vorstehend angeführten Aspekte und weitere Aspekte der Erfindung gehen aus dem nachfolgend beschrieben Ausführungsbeispiel hervor und sind anhand dieses Ausführungsbeispiels erläutert.

Die Erfindung wird im Folgenden anhand von einem in den Zeichnungen dargestellten Ausführungsbeispiel weiter beschrieben, auf das die Erfindung aber nicht beschränkt ist.

Die Figur 1 zeigt auf schematische Weise in Form eines Blockschaltbilds eine Spracherkennungseinrichtung gemäß einem Ausführungsbeispiel der Erfindung.

Die Figur 2 zeigt auf analoge Weise wie die Figur 1 Audio-Preprozessormittel der Spracherkennungsvorrichtung gemäß der Figur 1.

Die Figur 3 zeigt auf analoge Weise wie die Figur 1 Featurevektor-Extrahierungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 4 zeigt auf analoge Weise wie die Figur 1 Empfangskanal-Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 5 zeigt auf analoge Weise wie die Figur 1 erste Spracheigenschaft-Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 6 zeigt auf analoge Weise wie die Figur 1 zweite Spracheigenschaft-Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 7 zeigt auf analoge Weise wie die Figur 1 dritte Spracheigenschaft-Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 8 zeigt auf analoge Weise wie die Figur 1 vierte Spracheigenschaft-Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 9 zeigt auf analoge Weise wie die Figur 1 Sprach-Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 10 zeigt auf analoge schematische Weise in Form eines Diagramms einen zeitlichen Aktivitätsverlauf mehrerer Erkennungsmittel der Spracherkennungseinrichtung gemäß der Figur 1.

Die Figur 11 zeigt auf analoge Weise wie die Figur 1 ein Detail der Audio-Preprozessormittel gemäß der Figur 2.

Die Figur 12 zeigt auf analoge Weise wie die Figur 1 eine Logarithmus-Filterbank-Stufe der Featurevektor-Extrahierungsmittel gemäß der Figur 3.

Die Figur 13 zeigt auf analoge Weise wie die Figur 1 eine Musik-Erkennungsstufe der ersten Spracheigenschaft-Erkennungsmittel gemäß der Figur 5.

Die Figur 14 zeigt auf analoge Weise wie die Figur 1 eine zweite Trainingsstufe der zweiten Spracheigenschaft-Erkennungsmittel gemäß der Figur 6.

Die Figur 15 zeigt auf analoge Weise wie die Figur 1 eine vierte Trainingstufe der dritten Spracheigenschaft-Erkennungsmittel gemäß der Figur 7.

Die Figur 16 zeigt auf analoge Weise wie die Figur 1 eine sechste Trainingsstufe der vierten Spracheigenschaft-Erkennungsmittel gemäß der Figur 8.

In der Figur 1 ist eine Spracherkennungseinrichtung 1 dargestellt, die zum Erkennen einer zu einer Sprachinformation SI korrespondierenden Textinformation TI ausgebildet ist und die eine Konferenz-Transkriptionseinrichtung realisiert, mit deren Hilfe die bei einer Konferenz auftretende und von Konferenzteilnehmern beim Sprechen erzeugte Sprachinformation SI in die Textinformation TI transkribierbar ist.

Die Spracherkennungseinrichtung 1 ist mit Hilfe eines Computers 1A realisiert, wobei in der Figur 1 nur für die Spracherkennungseinrichtung 1 relevanten Funktionsgruppen dargestellt sind. Der Computer 1A weist eine in der Figur 1 nicht dargestellte Recheneinheit und einen internen Speicher 1B auf, wobei nachfolgend im Zusammenhang mit der Figur 1 nur auf die für die Spracherkennungseinrichtung 1 relevante Funktionalität des internen Speichers 1B näher eingegangen ist. Die Spracherkennungseinrichtung 1 nützt zum Erkennen der zu der Sprachinformation SI korrespondierenden Textinformation TI den internen Speicher 1B. Der Computer arbeitet ein Computerprogramm-Produkt ab, das direkt in den Speicher 1B des Computers 1A geladen werden kann und das Softwarecodeabschnitte aufweist.

Die Spracherkennungseinrichtung 1 weist Empfangsmittel 2 auf, die zum Empfangen einer Sprachinformation SI und zum Erzeugen und zum Abgeben von die Sprachinformation SI repräsentierenden Audiosignalen AS ausgebildet sind, wobei eine das Erkennen der Sprachinformation SI beeinflussende Bandbreite des Audiosignals AS von einem zum Empfangen der Sprachinformation SI verwendeten Empfangskanal bzw. Übertragungskanal abhängt. Die Empfangsmittel 2 weisen eine erste Empfangsstufe 3 auf, die einen ersten Empfangskanal realisiert und mit deren Hilfe über eine Vielzahl von Mikrofonen 4 die Sprachinformation SI empfangbar ist, wobei jedes Mikrofon 4 einem der in einem Konferenzraum befindlichen Konferenzteilnehmer zugeordnet ist, von dem die

Sprachginformation SI erzeugbar ist. Den Mikrofonen 4 ist eine in der Figur 1 nicht dargestellte sogenannte „Soundkarte" des Computers 1A zugeordnet, mit deren Hilfe die analogen Audiosignale AS in digitale Audiosignals AS umwandelbar sind. Die Empfangsmittel 2 weisen weiters eine zweite Empfangsstufe 5 auf, die einen zweiten Empfangskanal realisiert und mit deren Hilfe über eine Vielzahl von analogen Telefonleitungen die Sprachinformation SI empfangbar ist. Die Empfangsmittel 2 weisen weiters eine dritte Empfangsstufe 6 auf, die einen dritten Empfangskanal realisiert und mit deren Hilfe über eine Vielzahl von ISDN-Telefonleitungen die Sprachinformation SI empfangbar ist. Die Empfangsmittel 2 weisen eine vierte Empfangsstufe 7 auf, die einen vierten Empfangskanal realisiert und mit deren Hilfe über ein Computer-Datennetzwerk die Sprachinformation SI mit Hilfe eines sogenannte „Voice-over-IP" Datenstroms empfangbar ist. Die Empfangsmittel 2 sind weiters zum Abgeben einer digitalen Repräsentation des empfangenen Audiosignals AS in Form eines Datenstroms ausgebildet, wobei die digitale Repräsentation des Audiosignals AS eine dem jeweiligen Empfangskanal entsprechende Audiosignalformatierung aufweist und wobei der Datenstrom sogenannte Audioblöcke und in den Audioblöcken enthaltene sogenannte Audioheader aufweist, welche Audioheader die jeweilige Audiosignalformatierung angeben.

Die Spracherkennungseinrichtung 1 weist weiters Audio-Preprozessormittel 8 auf, die zum Empfangen des von den Empfangsmitteln 2 abgegebenen Audiosignals AS ausgebildet sind. Die Audio-Preprozessormittel 8 sind weiters zum Umwandeln des empfangenen Audiosignals AS in ein für ein weiteres Verarbeiten vorgesehenes einheitlich formatiertes, nämlich einheitlich PCM-formatiertes Audiosignal PAS und zum Abgeben des Audiosignals PAS ausgebildet. Zu diesem Zwecke weisen die in der Figur 2 dargestellten Audio-Preprozessormittel 8 eine Kodierung-Erkennungsstufe 9, eine erste Datenstromsteuerstufe 10, eine Dekodierstufe 11, eine Dekodieralgorithmus-Auswahlstufe 12, eine Dekodieralgorithmus-Speicherstufe 13 und eine Hochpassfilterstufe 14 auf. Der ersten Datenstromsteuerstufe 10 ist das empfangene Audiosignal AS direkt zuführbar. Der Kodierung-Erkennungsstufe 9 sind die Audioheader zuführbar. Die Kodierung-Erkennungsstufe 9 ist an Hand der Audioheader zum Erkennen einer möglichen Kodierung des durch die Audioblöcke repräsentierten Audiosignals AS und bei Vorliegen einer Kodierung zum Abgeben einer Kodierung-Erkennungsinformation COI an die

Dekodieralgorithmus-Auswahlstufe 12 ausgebildet. Weiters ist die Kodierung-Erkennungsstufe 9 bei Vorliegen einer Kodierung zum Abgeben einer Datenstrom-Beeinflussungsinformation DCSI an die ersten Datenstromsteuerstufe 10 ausgebildet, so dass das der ersten Datenstromsteuerstufe 10 zugeführte Audiosignal AS an die Dekodierstufe 11 abgebbar ist. Bei einem Nichtfeststellen einer Kodierung des Audiosignals AS ist von der Kodierung-Erkennungsstufe 9 mit Hilfe der Datenstrom-Beeinflussungsinformation DCSI die Datenstromsteuerstufe 10 derart steuerbar, dass das Audiosignal AS von der Datenstromsteuerstufe 10 direkt an die Hochpassfilterstufe 14 abgebbar ist.

Die Dekodieralgorithmus-Speicherstufe 13 ist zum Speichern einer Vielzahl von Dekodieralgorithmen ausgebildet. Die Dekodieralgorithmus-Auswahlstufe 12 ist durch ein Softwareobjekt realisiert, das in Abhängigkeit von der Kodierung-Erkennungsinformation COI zum Auswählen von einem der gespeicherten Dekodieralgorithmen und unter Ausnutzung des gewählten Dekodieralgorithmus zum Erzeugen der Dekodierstufe 11 ausgebildet. Die Dekodierstufe 11 ist in Abhängigkeit von dem ausgewählten Decodieralgorithmus zum Dekodieren des Audiosignals AS und zum Abgeben eines kodierungsfreien Audiosignals AS an die Hochpassfilterstufe 14 ausgebildet. Die Hochpassfilterstufe 14 ist zum Hochpassfiltern des Audiosignals AS ausgebildet, so dass störende niederfrequente Anteile des Audiosignals AS entfernbar sind, welche niederfrequenten Anteile eine weitere Verarbeitung des Audiosignals AS nachteilig beeinflussen können.

Die Audio-Preprozessormittel 8 weisen weiters eine PCM-Format-Umwandlungsparameter-Erzeugungsstufe 15, die zum Empfangen des hochpassgefilterten Audiosignals AS und zum Verarbeiten von einer zu dem hochpassgefilterten Audiosignal AS gehörenden PCM-Format-Information PCMF ausgebildet ist, wobei die PCM-Format-Information PCMF von dem jeweiligen Audioheader repräsentiert ist. Die PCM-Format-Umwandlungsparameter-Erzeugungsstufe 15 ist weiters unter Ausnutzung der PCM-Format-Information PCMF und unter Ausnutzung einer in der Figur 2 nicht dargestellten definierbaren PCM-Format-Konfigurationsinformation PCMC, die das zu erzeugenden einheitliche PCM-Format des Audiosignals PAS angibt, zum Erzeugen und zum Abgeben von PCM-Format-Umwandlungsparametern PCP ausgebildet.

Die Audio-Preprozessormittel 8 weisen weiters eine Umwandlungsstufen-

Erzeugungsstufe 16 auf, die durch ein Softwareobjekt realisiert ist und die zum Empfangen und zum Verarbeiten der PCM-Format-Umwandlungsparameter PCP und unter Ausnutzung dieser Parameter PCP zum Erzeugen einer PCM-Format-Umwandlungsstufe 17 ausgebildet sind. Die PCM-Format-Umwandlungsstufe 17 ist zum Empfangen des hochpassgefilterten Audiosignals AS und zum Umwandeln des hochpassgefilterten Audiosignals AS in das Audiosignal PAS und zum Abgeben des Audiosignals PAS von den Audio-Preprozessormitteln 8 ausgebildet. Die PCM-Format-Umwandlungsstufe 17 weist – in der Figur 2 nicht dargestellte - eine Vielzahl von in Abhängigkeit von den PCM-Format-Umwandlungsparametern PCP erzeugbaren Umwandlungsstufen zum Realisieren der PCM-Format-Umwandlungsstufe 17.

Die in der Figur 11 im Detail dargestellte PCM-Format-Umwandlungsparameter-Erzeugungsstufe 15 weist eingangsseitig eine Parser-Stufe 15A auf, die unter Ausnutzung der PCM-Format-Konfigurationsinformation PCMC und der PCM-Format-Information PCMF zum Bestimmen der Anzahl der Umwandlungsstufen der Format-Umwandlungsstufen 17 und der ihnen individuell zugeordneten Eingang/Ausgang-PCM-Formate ausgebildet ist, was durch eine von ihr abgebbare Objekt-Spezifikationsinformation OSI repräsentiert ist. Dabei definiert die PCM-Format-Information PCMF ein Eingang-Audiosignalformat und die PCM-Format-Konfigurationsinformation PCMC ein Ausgang-Audiosignalformat der PCM-Format-Umwandlungsparameter-Erzeugungsstufe 15. Die PCM-Format-Umwandlungsparameter-Erzeugungsstufe 15 weist weiters eine Filterplanerstufe 15B auf, die unter Ausnutzung der Objekt-Spezifikationsinformation OSI zum Planen weiterer Eigenschaften jeder der Umwandlungsstufen ausgebildet ist, welche weiteren Eigenschaften und die Objekt-Spezifikationsinformation OSI durch die von ihr erzeugbare und abgebbare PCM-Format-Umwandlungsparameter PCP repräsentiert sind.

Die in der Figur 1 dargestellte Spracherkennungseinrichtung 1 weist weiters Empfangskanal-Erkennungsmittel 18 auf, die zum Empfangen des von den Audio-Preprozessormitteln 8 vorverarbeiteten Audiosignals PAS und zum Erkennen des jeweils zum Empfangen der Sprachinformation SI verwendeten Empfangskanals und zum Erzeugen einer den erkannten Empfangskanal repräsentierenden Kanalangabe-Information CHI und zum Abgeben der Kanalangabe-Information CHI ausgebildet sind.

Die Spracherkennungseinrichtung 1 weist weiters Featurevektor-

Extrahierungsmittel 19 auf, die ebenfalls wie die Empfangskanal-Erkennungsmittel 18 zum Empfangen des durch die Audio-Preprozessormittel 8 vorverarbeiteten Audiosignals PAS und der Kanalangabe-Information CHI und unter Berücksichtigung der Kanalangabe-Information CHI zum Erzeugen und zum Abgeben von sogenannten Featurevektoren FV ausgebildet sind, worauf an geeigneter Stelle im Zusammenhang mit der Figur 3 noch im Detail eingegangen wird.

Die Spracherkennungseinrichtung 1 weist weiters erste Spracheigenschaft-Erkennungsmittel 20 auf, die zum Empfangen der die Sprachinformation SI repräsentierenden Featurevektoren FV und zum Empfangen der Kanalangabe-Information CHI ausgebildet sind. Die ersten Spracheigenschaft-Erkennungsmittel 20 sind weiters unter Ausnutzung der Featurevektoren FV und unter fortwährender Berücksichtigung der Kanalangabe-Information CHI zum Erkennen einer ersten Spracheigenschaft - nämlich einer akustischen Segmentierung - und zum Erzeugen und zum Abgeben einer die erkannte akustische Segmentierung repräsentierenden ersten Eigenschaftsinformation - nämlich einer Segmentierung-Information ASI - ausgebildet.

Die Spracherkennungseinrichtung 1 weist weiters zweite Spracheigenschaft-Erkennungsmittel 21 auf, die zum Empfangen der die Sprachinformation SI repräsentierenden Featurevektoren FV und zum Empfangen der Kanalangabe-Information CHI und zum Empfangen der Segmentierung-Information ASI ausgebildet sind. Die zweiten Spracheigenschaft-Erkennungsmittel 21 sind weiters unter Ausnutzung der Featurevektoren FV und unter fortwährender Berücksichtigung der Kanalangabe-Information CHI und der Segmentierung-Information ASI zum Erkennen einer zweiten Spracheigenschaft - nämlich um welche Sprache es sich handelt, also beispielsweise Englisch oder Französisch oder Spanisch - und zum Erzeugen und zum Abgeben einer die erkannte Sprache repräsentierenden zweiten Eigenschaftsinformation - nämlich einer Sprache-Information LI - ausgebildet.

Die Spracherkennungseinrichtung 1 weist weiters dritte Spracheigenschaft-Erkennungsmittel 22 auf, die zum Empfangen der die Sprachinformation SI repräsentierenden Featurevektoren FV, der Kanalangabe-Information CHI, der Segmentierung-Information ASI und der Sprache-Information LI ausgebildet sind. Die dritten Spracheigenschaft-Erkennungsmittel 22 sind weiters unter Ausnutzung der Featurevektoren FV und unter fortwährender Berücksichtigung der Informationen CHI,

ASI und LI zum Erkennen einer dritten Spracheigenschaft - nämlich einer Sprechergruppe - und zum Erzeugen und zum Abgeben einer die erkannte Sprechergruppe repräsentierenden dritten Eigenschaftsinformation - nämlich einer Sprechergruppe-Information SGI - ausgebildet.

Die Spracherkennungseinrichtung 1 weist weiters vierte Spracheigenschaft-Erkennungsmittel 23 auf, die zum Empfangen der die Sprachinformation SI repräsentierenden Featurevektoren FV und zum Empfangen der Kanalangabe-Information CHI, der Segmentierung-Information ASI, der Sprache-Information LI und der Sprechergruppe-Information SGI ausgebildet sind. Die vierten Spracheigenschaft-Erkennungsmittel 23 sind weiters unter Ausnutzung der Featurevektoren FV und unter fortwährender Berücksichtigung der Informationen CHI, ASI, LI und SGI zum Erkennen einer vierten Spracheigenschaft - nämlich eines Kontexts - und zum Erzeugen und zum Abgeben einer den erkannten Kontext repräsentierenden vierten Eigenschaftsinformation - nämlich einer Kontext-Information CI - ausgebildet.

Die Spracherkennungseinrichtung 1 weist weiters Sprach-Erkennungsmittel 24 auf, die unter fortwährender Berücksichtigung der Kanalangabe-Information CHI, der ersten Eigenschaftsinformation ASI, der zweiten Eigenschaftsinformation LI, der dritten Eigenschaftsinformation SGI und der vierten Eigenschaftsinformation CI zum Erkennen der Textinformation TI unter Ausnutzung der die Sprachinformation SI repräsentierenden Featurevektoren FV und zum Abgeben der Textinformation TI ausgebildet sind.

Die Spracherkennungseinrichtung 1 weist weiters Textinformation-Speichermittel 25 und Textinformation-Bearbeitungsmittel 26 und Textinformation-Ausgabemittel 27 auf, wobei die Mittel 25 und 27 zum Empfangen der Textinformation TI von den Sprach-Erkennungsmitteln 24 her ausgebildet sind. Die Textinformation-Speichermittel 25 sind zum Speichern der Textinformation TI und zum Bereitstellen der Textinformation TI für ein weiteres Verarbeiten mit Hilfe der Mittel 26 und 27 ausgebildet.

Die Textinformation-Bearbeitungsmittel 26 sind zum Zugreifen auf die in den Textinformation-Speichermitteln 25 gespeicherte Textinformation TI und zum Bearbeiten der durch die Sprach-Erkennungsmittel 24 automatisch aus der Sprachinformation SI erzeugbaren Textinformation TI ausgebildet. Zu diesem Zweck weisen die Textinformation-Bearbeitungsmittel 26 in der Figur 1 nicht dargestellte Anzeige/Eingabe-Mittel auf, die es einem Benutzer – beispielsweise einer Korrektionistin – erlauben, die

Textinformation TI zu bearbeiten, so dass bedingt durch eine undeutliche oder fehlerhafte Aussprache eines Konferenzteilnehmers oder durch Probleme bei der Übertragung des Audiosignals AS bei dem automatischen Transkribieren verursachte Unklarheiten oder Fehler in der Textinformation TI auf manuelle Weise bereinigbar sind.

Die Textinformation-Ausgabemittel 27 sind zum Ausgeben der in den Textinformation-Speichermitteln 25 gespeicherten und gegebenenfalls durch einen Benutzer bearbeiteten Textinformation TI ausgebildet, wobei die Textinformation-Ausgabemittel 27 in der Figur 1 nicht dargestellte Schnittstellenmittel zum Abgeben der Textinformation TI in Form eines digitalen Datenstroms an ein Computernetzwerk, an eine Druckvorrichtung und an eine Anzeigevorrichtung aufweisen.

Im Nachfolgenden soll ein zeitliches Zusammenwirken der Erkennungsmittel 18, 20, 21, 22, 23 und 24 an Hand eines zeitlichen Aktivitätsverlaufs der Erkennungsmittel 18, 20, 21, 22, 23 und 24 mit Hilfe der Figur 10 erläutert werden. Zu diesem Zweck sind in der Figur 10 die einzelnen Aktivitäten in Form eines Balkendiagramms dargestellt, wobei ein erster Aktivitätsbalken 28 die Aktivität der Empfangskanal-Erkennungsmittel 18 repräsentiert und wobei ein zweiter Aktivitätsbalken 29 die Aktivität der ersten Spracheigenschaft-Erkennungsmittel 20 repräsentiert und wobei ein dritter Aktivitätsbalken 30 die Aktivität der zweiten Spracheigenschaft-Erkennungsmittel 21 repräsentiert und wobei ein vierter Aktivitätsbalken 31 die Aktivität der dritten Spracheigenschaft-Erkennungsmittel 22 repräsentiert und wobei ein fünfter Aktivitätsbalken 32 die Aktivität der vierten Spracheigenschaft-Erkennungsmittel 23 repräsentiert und wobei ein sechster Aktivitätsbalken 33 die Aktivität der Sprach-Erkennungsmittel 24 repräsentiert.

Der erste Aktivitätsbalken 28 erstreckt sich von einem ersten Startzeitpunkt T1B bis zu einem ersten Endzeitpunkt T1E. Der zweite Aktivitätsbalken 29 erstreckt sich von einem zweiten Startzeitpunkt T2B bis zu einem zweiten Endzeitpunkt T2E. Der dritte Aktivitätsbalken 30 erstreckt sich von einem dritten Startzeitpunkt T3B bis zu einem dritten Endzeitpunkt T3E. Der vierte Aktivitätsbalken 31 erstreckt sich von einem vierten Startzeitpunkt T4B bis zu einem vierten Endzeitpunkt T4E. Der fünfte Aktivitätsbalken 32 erstreckt sich von einem fünften Startzeitpunkt T5B bis zu einem fünften Endzeitpunkt T5E. Der sechste Aktivitätsbalken 33 erstreckt sich von einem sechsten Startzeitpunkt T6B bis zu einem sechsten Endzeitpunkt T6E. Dabei wird während der Aktivität des jeweiligen

Erkennungsmittels 18, 20, 21, 22, 23 oder 24 von dem jeweiligen Erkennungsmittel 18, 20, 21, 22, 23 oder 24 die gesamte Sprachinformation SI vollständig verarbeitet, wobei jedes der Erkennungsmittel 18, 20, 21, 22, 23 und 24 das Verarbeiten der Sprachinformation SI beginnend am Anfang der Sprachinformation SI zu dem jeweiligen ihm zugeordneten Startzeitpunkt T1B, T2B, T3B, T4B, T5B bzw. T6B beginnt und zu dem jeweiligen ihm zugeordneten Endzeitpunkt T1E, T2E, T3E, T4E, T5E bzw. T6E beendet. Üblicherweise unterscheiden sich die zwischen den Startzeitpunkten T1B, T2B, T3B, T4B, T5B bzw. T6B und den Endzeitpunkten T1E, T2E, T3E, T4E, T5E bzw. T6E vorliegenden Gesamtverarbeitungszeitspannen praktisch nicht voneinander. Es können jedoch Unterschiede bei den individuellen Gesamtverarbeitungszeitspannen auftreten, wenn die jeweiligen Verarbeitungsgeschwindigkeiten der Mittel 18, 20, 21, 22, 23 und 24 voneinander abweichen, was beispielsweise dann zum Tragen kommt, wenn die Sprachinformation SI „off-line-mäßig" verfügbar gemacht wird. Dabei ist unter dem Begriff „off-line-mäßig" beispielsweise eine vorangehende Aufzeichnung der Sprachinformation SI auf einem Aufzeichnungsträger zu verstehen, der nachfolgend daran der Spracherkennungseinrichtung 1 zugänglich gemacht wird.

Weiters sind in dem Diagramm zu den jeweiligen Erkennungsmitteln 18, 20, 21, 22, 23 und 24 korrespondierende Startverzögerungen d1 bis d6 dargestellt, wobei im vorliegenden Fall d1=0 ist, weil der Nullpunkt der Zeitachse T zeitlich zusammenfallend mit dem ersten Startzeitpunkt T1B der Empfangskanal-Erkennungsmittel 18 gewählt wurde. Es sei jedoch erwähnt, dass dieser Nullpunkt auch zu einem anderen Zeitpunkt gewählt werden kann, wodurch d1 ungleich Null wird.

Weiters sind in dem Diagramm zu den jeweiligen Erkennungsmitteln 18, 20, 21, 22, 23 und 24 korrespondierende anfängliche Verarbeitungsverzögerungen D1 bis D6 eingetragen, die durch das jeweilige Erkennungsmittel 18, 20, 21, 22, 23 und 24 bei einem erstmaligen Erzeugen der jeweiligen Information CHI, ASI, LI, SGI, CI bzw. TI selbst verursacht sind. Mathematisch lässt sich der Zusammenhang zwischen d_i und D_i wie folgt zusammenfassen, wobei per Definition $d_0 = 0$ und $D_0 = 0$ ist:

$$d_i = d_{i-1} + D_{i-1} \quad i = 1 \ldots 6 \text{ und daraus folgend:}$$

$$d_i = \sum_{i=0}^{i-1} D_i + d_0 \quad i = 1 \ldots 6.$$

Die Empfangskanal-Erkennungsmittel 18 beginnen zu dem ersten

Startzeitpunkt T1B mit dem Erkennen des jeweils zum Empfangen der Sprachinformation SI verwendeten Empfangskanals 3, 5, 6 oder 7. Dabei erfolgt das Erkennen des jeweiligen Empfangskanals 3, 5, 6 oder 7 während einer ersten anfänglichen Verarbeitungsverzögerung D1 für einen Teilbereich eines ersten Teils der Sprachinformation SI, welcher erste Teil während der Verarbeitungsverzögerung D1 von den Audio-Preprozessormitteln 8 vorverarbeitet an die Empfangskanal-Erkennungsmittel 18 abgebbar ist und welcher erste Teil während der Verarbeitungsverzögerung D1 von den Empfangskanal-Erkennungsmitteln 18 zum erstmaligen Erkennen des verwendeten Empfangskanals 3, 5, 6 oder 7 ausnutzbar ist. Im vorliegenden Fall beträgt die Verarbeitungsverzögerung D1 etwa einhundert (100) Millisekunden und der erste Teil der Sprachinformation SI umfasst etwa zehn (10) sogenannte Frames, wobei jeder Frame die Sprachinformation SI während einer Zeitdauer von etwa zehn Millisekunden in der Audiosignalebene repräsentiert. Die Empfangskanal-Erkennungsmittel 18 erzeugen am Ende der Verarbeitungsverzögerung D1 erstmalig die den erkannten Empfangskanal 3, 5, 6 oder 7 repräsentierende Kanalangabeinformation CHI für einen ersten Frame des ersten Teils der Sprachinformation SI und geben diese Kanalangabe-Information CHI an die vier Spracheigenschaft-Erkennungsmittel 20 bis 23 und an die Sprach-Erkennungsmittel 24 ab. In dem Diagramm ist dies mit Hilfe des Pfeilbüschels 34 angedeutet.

Im weiteren Zeitverlauf bis hin zu dem Endzeitpunkt T1E erzeugen bzw. stellen die Empfangskanal-Erkennungsmittel 18 fortwährend eine frameweise aktualisierte Kanalangabe-Information CHI für die vier Spracheigenschaft-Erkennungsmittel 20 bis 23 und die Sprach-Erkennungsmittel 24 zur Verfügung, so dass die Kanalangabe-Information CHI fortwährend frameweise von den Erkennungsmitteln 20 bis 24 berücksichtigbar ist. Dabei wird beginnend mit dem zweiten Frame der Sprachinformation SI jeweils ein weiterer Teil der Sprachinformation SI verarbeitet, der eine den Umständen angepasste Anzahl von Frames aufweist, und eine jeweils für den ersten Frame, also für den ersten Teilbereich des jeweiligen Teils der Sprachinformation SI gültige Kanalangabe-Information CHI erzeugt bzw. bereitgestellt. Dabei unterscheiden sich benachbarte Teile der Sprachinformation SI, beispielsweise der erste Teil und ein zweiter Teil, dahingehend, dass der zweite Teil als einen letzten Frame einen an den ersten Teil angrenzenden Frame aufweist, der jedoch nicht in dem ersten Teil enthalten ist, und dass der erste Frame des zweiten Teils durch einen an den ersten Frame des ersten Teils anschließenden zweiten

Frame des ersten Teils gebildet ist.

Es sei an dieser Stelle erwähnt, dass nach dem erstmaligen Erzeugen bei dem weiteren, also fortwährenden Erzeugen der Kanalangabe-Information CHI in Abhängigkeit von dem Auftreten des Audiosignals AS bei einem der Empfangskanäle 3, 5, 6 und 7 auch andere Zeitspannen als die erste anfängliche Verarbeitungsverzögerung D1 auftreten können und demgemäß auch eine andere Anzahl von Frames zum Erzeugen der Kanalangabe-Information CHI für den ersten Frame der jeweiligen Anzahl von Frames, also für den ersten Frame der weiteren Teile der Sprachinformation SI berücksichtigbar ist. Es sei an dieser Stelle weiters erwähnt, dass sich benachbarte Teile der Sprachinformation SI auch um mehr als zwei Frames unterscheiden können. Weiters sei erwähnt, dass der Teilbereich eines Teils der Sprachinformation SI, für den die Kanalangabe-Information CHI erzeugt wird, auch mehrere Frames umfassen kann, wobei diese mehreren Frames bevorzugt am Anfang eines Teils der Sprachinformation SI lokalisiert sind. Weiters sei erwähnt, dass der jeweilige Teilbereich eines Teils der Sprachinformation SI, für den die Kanalangabe-Information CHI erzeugt wird, auch die Gesamtanzahl der Frames des Teils der Sprachinformation SI aufweisen kann, so dass der Teilbereich identisch zu dem Teil ist. Es sei weiters erwähnt, dass der jeweilige Teilbereich eines Teils der Sprachinformation SI, für den die Kanalangabe-Information CHI erzeugt wird, nicht unbedingt der erste Frame, sondern auch der zweite Frame oder jeder weitere Frame des Teils der Sprachinformation SI sein kann. Dabei ist wichtig zu verstehen, dass zu einem Frame genau eine einzige Kanalangabe-Information CHI zugeordnet ist.

Vorwegnehmend sei an dieser Stelle festgehalten, dass die vorstehend gemachten Angaben hinsichtlich eines Teils des Sprachsignals SI und hinsichtlich des Teilbereichs des jeweiligen Teils der Sprachinformation SI, für den die jeweilige Information ASI, LI, SGI, CI und TI erzeugt wird, auch bei den nachfolgend beschriebenen Mitteln 20, 21, 22, 23 und 24 Gültigkeit haben.

Beginnend zu dem Zeitpunkt T2B beginnen die ersten Spracheigenschaft-Erkennungsmittel 20 um die Startverzögerung d2 zeitverzögert unter Ausnutzung der den ersten Teil der Sprachinformation SI repräsentierenden Featurevektoren FV und unter Berücksichtigung der zu jedem Frame des ersten Teils der Sprachinformation SI jeweils zugeordneten Kanalangabe-Information CHI mit dem erstmaligen Erkennen der akustischen Segmentierung für den ersten Frame, also für den ersten Teilbereich des ersten

Teils der Sprachinformation SI. Die Startverzögerung d2 entspricht dabei der durch die Empfangskanal-Erkennungsmittel 18 verursachten anfänglichen Verarbeitungsverzögerung D1. Demgemäß sind die ersten Spracheigenschaft-Erkennungsmittel 20 zeitverzögert um mindestens die Zeitspanne, die von den Empfangskanal-Erkennungsmitteln 18 zum Erzeugen der Kanalangabe-Information CHI für den ersten Frame benötigt wird, zum erstmaligen Erkennen der akustischen Segmentierung für den ersten Frame ausgebildet. Auch die ersten Spracheigenschaft-Erkennungsmittel 20 weisen ihrerseits eine zweite anfängliche Verarbeitungsverzögerung D2 auf, wobei nach Verstreichen dieser Verarbeitungsverzögerung D2 erstmals die Segmentierung-Information ASI für den ersten Frame des ersten Teils der Sprachinformation SI erzeugbar und an die Erkennungsmittel 21 bis 24 abgebbar ist, was stellvertretend für ein weiteres in der Figur 11 nicht dargestelltes Pfeilbüschel durch einen einzigen Pfeil 35 angedeutet ist.

Nachfolgend an die Verarbeitungsverzögerung D2 wird von den ersten Spracheigenschaft-Erkennungsmitteln 20 unter fortwährender Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondiererenden Kanalangabe-Information CHI fortwährend für die nach dem ersten Frame der Sprachinformation SI auftretenden weiteren Frames, nämlich für jeden ersten Frame des jeweiligen Teils der Sprachinformation SI, eine aktualisierte Segmentierung-Information ASI erzeugt bzw. bereitgestellt.

Beginnend zu dem Zeitpunkt T3B beginnen die zweiten Spracheigenschaft-Erkennungsmittel 21 um die Startverzögerung d3 zeitverzögert unter Ausnutzung der den ersten Teil der Sprachinformation SI repräsentierenden Featurevektoren FV und unter Berücksichtigung der zu jedem Frame des ersten Teils der Sprachinformation SI jeweils zugeordnete Kanalangabe-Information CHI und der Segmentierung-Information ASI mit dem erstmaligen Erkennen der Sprache für den ersten Frame, also für den ersten Teilbereich des ersten Teils der Sprachinformation SI. Die Startverzögerung d3 entspricht dabei der durch die Empfangskanal-Erkennungsmitteln 18 und die ersten Spracheigenschaft-Erkennungsmittel 20 verursachten Summe der anfänglichen Verarbeitungsverzögerungen D1 und D2. Demgemäß sind die zweiten Spracheigenschaft-Erkennungsmittel 20 zeitverzögert um mindestens die Zeitspanne, die von den Empfangskanal-Erkennungsmitteln 18 und den Spracheigenschaft-Erkennungsmitteln 20 zum erstmaligen Erzeugen der Kanalangabe-Information CHI und der Segmentierung-

Information ASI für den ersten Frame benötigt werden, zum erstmaligen Erkennen der Sprache für den ersten Frame ausgebildet. Auch die zweiten Spracheigenschaft-Erkennungsmittel 21 weisen ihrerseits eine dritte anfängliche Verarbeitungsverzögerung D3 auf, wobei nach Verstreichen dieser Verarbeitungsverzögerung D3 erstmals die Sprache-Information ASI für den ersten Frame der Sprachinformation SI an die Erkennungsmittel 22 bis 24 erzeugbar und abgebbar ist, was stellvertretend für ein weiteres in der Figur 11 nicht dargestelltes Pfeilbüschel durch den einzigen Pfeil 36 angedeutet ist.

Nachfolgend an die Verarbeitungsverzögerung D3 wird von den zweiten Spracheigenschaft-Erkennungsmitteln 21 unter fortwährender Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Informationen CHI und ASI fortwährend für die nach dem ersten Frame der Sprachinformation SI auftretenden weiteren Frames, nämlich für jeden ersten Frame des jeweiligen Teils der Sprachinformation SI, eine aktualisierte Sprache-Information LI erzeugt bzw. bereitgestellt.

Beginnend zu dem Zeitpunkt T4B beginnen die dritten Spracheigenschaft-Erkennungsmittel 22 um die Startverzögerung d4 zeitverzögert unter Ausnutzung der den ersten Teil der Sprachinformation SI repräsentierenden Featurevektoren FV und unter Berücksichtigung der zu jedem Frame des ersten Teils der Sprachinformation SI jeweils zugeordneten Kanalangabe-Information CHI und Segmentierung-Information ASI und Sprache-Information LI mit dem erstmaligen Erkennen der Sprechergruppe für den ersten Frame, also für den ersten Teilbereich des ersten Teils der Sprachinformation SI. Die Startverzögerung d4 entspricht dabei der durch die Empfangskanal-Erkennungsmittel 18 und die ersten Spracheigenschaft-Erkennungsmittel 20 und die zweiten Spracheigenschaft-Erkennungsmittel 21 verursachten Summe der anfänglichen Verarbeitungsverzögerungen D1 und D2 und D3. Demgemäß sind die dritten Spracheigenschaft-Erkennungsmittel 22 zeitverzögert um mindestens die Zeitspanne, die von den Mitteln 18, 20 und 21 zum erstmaligen Erzeugen der Kanalangabe-Information CHI und der Segmentierung-Information ASI und der Sprache-Information LI für den ersten Frame benötigt werden, zum erstmaligen Erkennen der Sprechergruppe für den ersten Frame ausgebildet. Auch die dritten Spracheigenschaft-Erkennungsmittel 22 weisen ihrerseits eine vierte anfängliche Verarbeitungsverzögerung D4 auf, wobei nach Verstreichen dieser Verarbeitungsverzögerung D4 erstmals die Sprechergruppe-Information SGI für den ersten

Frame an die Erkennungsmittel 23 und 24 erzeugbar und abgebbar ist, was stellvertretend für ein weiteres in der Figur 11 nicht dargestelltes Pfeilbüschel durch einen einzigen Pfeil 37 angedeutet ist.

Nachfolgend an die Verarbeitungsverzögerung D4 wird von den dritten Spracheigenschaft-Erkennungsmitteln 22 unter fortwährender Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Information CHI, ASI und LI fortwährend für die nach dem ersten Frame der Sprachinformation SI auftretenden weiteren Frames, nämlich für jeden ersten Frame des jeweiligen Teils der Sprachinformation SI, eine aktualisierte Sprechergruppe-Information SGI erzeugt bzw. bereitgestellt.

Beginnend zu dem Zeitpunkt T5B beginnen die vierten Spracheigenschaft-Erkennungsmittel 23 um die Startverzögerung d5 zeitverzögert unter Ausnutzung der den ersten Teil der Sprachinformation SI repräsentierenden Featurevektoren FV und unter Berücksichtigung der zu jedem Frame des ersten Teils der Sprachinformation SI jeweils zugeordneten Kanalangabe-Information CHI und Segmentierung-Information ASI und Sprache-Information LI und Sprechergruppe-Information SGI mit dem erstmaligen Erkennen des Kontexts für den ersten Frame, also für den ersten Teilbereich des ersten Teils der Sprachinformation SI. Die Startverzögerung d5 entspricht dabei der durch die Mittel 18, 20, 21 und 22 verursachten Summe der anfänglichen Verarbeitungsverzögerungen D1 und D2 und D3 und D4. Demgemäß sind die vierten Spracheigenschaft-Erkennungsmittel 23 zeitverzögert um mindestens die Zeitspannen, die von den Mitteln 18, 20, 21 und 22 zum erstmaligen Erzeugen der Informationen CHI, ASI, LI und SGI für den ersten Frame benötigt werden, zum erstmaligen Erkennen des Kontexts für den ersten Frame ausgebildet. Auch die vierten Spracheigenschaft-Erkennungsmittel 23 weisen ihrerseits eine fünfte anfängliche Verarbeitungsverzögerung D5 auf, wobei nach Verstreichen dieser Verarbeitungsverzögerung D5 erstmals die Kontext-Information CI für den ersten Frame der Sprachinformation SI an die Sprach-Erkennungsmittel 24 erzeugbar und abgebbar ist, was durch einen Pfeil 38 angedeutet ist.

Nachfolgend an die Verarbeitungsverzögerung D5 wird von den vierten Spracheigenschaft-Erkennungsmitteln 23 unter fortwährender Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Informationen CHI, ASI, LI und SGI fortwährend für die nach dem ersten Frame der

Sprachinformation SI auftretenden weiteren Frames, nämlich für jeden ersten Frame des jeweiligen Teils der Sprachinformation SI, eine aktualisierte Kontext-Information CI erzeugt bzw. bereitgestellt.

Beginnend zu dem Zeitpunkt T6B beginnen die Sprach-Erkennungsmittel 24 um die Startverzögerung d6 zeitverzögert unter Ausnutzung der den ersten Teil der Sprachinformation SI repräsentierenden Featurevektoren FV und unter Berücksichtigung der zu jedem Frame des ersten Teils der Sprachinformation SI jeweils zugeordneten Kanalangabe-Information CHI und Segmentierung-Information ASI und Sprache-Informationen LI und Sprechergruppe-Information SGI und Kontext-Information CI mit dem erstmaligen Erkennen der Textinformation TI für den ersten Frame, also für den ersten Teilbereich des ersten Teils der Sprachinformation SI. Die Startverzögerung d6 entspricht dabei der durch die Mittel 18, 20, 21, 22 und 23 verursachten Summe der anfänglichen Verarbeitungsverzögerungen D1 und D2 und D3 und D4 und D5. Demgemäß sind die Erkennungsmittel 24 zeitverzögert um mindestens die Zeitspannen, die von den Mitteln 18, 20, 21, 22 und 23 zum erstmaligen Erzeugen der Informationen CHI, ASI, LI, SGI und CI für den ersten Frame benötigt werden, zum erstmaligen Erkennen der Textinformation TI für den ersten Frame der Sprachinformation SI ausgebildet. Auch die Sprach-Erkennungsmittel 24 weisen ihrerseits eine anfängliche Verarbeitungsverzögerung D6 auf, wobei nach dem Verstreichen dieser Verarbeitungsverzögerung D6 erstmals die Textinformation TI für den ersten Frame der Sprachinformation SI an die Mittel 25, 26 bzw. 27 erzeugbar und abgebbar ist.

Nachfolgend an die Verarbeitungsverzögerung D6 wird von den Sprach-Erkennungsmitteln 24 unter fortwährender Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Informationen CHI, ASI, LI, SGI und CI fortwährend für die nach dem ersten Frame der Sprachinformation SI auftretenden weiteren Frames, nämlich für jeden ersten Frame des jeweiligen Teils der Sprachinformation SI, eine aktualisierte Textinformation TI erzeugt bzw. bereitgestellt.

Zusammenfassend sei im Zusammenhang mit den zeitlichen Aktivitäten erwähnt, dass immer dann ein Frame von einer der Erkennungsstufen 20, 21, 22, 23 oder 24 verarbeitet wird, wenn alle von der jeweiligen Erkennungsstufe 20, 21, 22, 23 oder 24 zum Verarbeiten des jeweiligen Frames benötigten Informationen CHI, ASI, LI, SGI bzw. CI bei der jeweiligen Erkennungsstufe 20, 21, 22, 23 oder 24 verfügbar sind.

Gemäß den vorstehenden Erläuterungen ist die Spracherkennungseinrichtung 1 zum Durchführen eines Spracherkennungsverfahrens zum Erkennen der zu der Sprachinformation SI korrespondierenden Textinformation TI ausgebildet, wobei die Sprachinformation SI hinsichtlich ihrer Spracheigenschaften, nämlich der akustischen Segmentierung, der Sprache, der Sprecher-Gruppe und des Kontexts charakterisierbar ist. Das Spracherkennungsverfahren weist die nachfolgend angeführten Verfahrensschritte auf, nämlich Erkennen der akustischen Segmentierung unter Ausnutzung der Sprachinformation SI und Erzeugen der die erkannte akustische Segmentierung repräsentierenden Segmentierung-Information ASI und Erkennen der Sprache unter Ausnutzung der Sprachinformation SI und Erzeugen der die erkannte Sprache repräsentierenden Sprache-Information LI und Erkennen der Sprecher-Gruppe unter Ausnutzung der Sprachinformation SI und Erzeugen der die erkannte Sprecher-Gruppe repräsentierenden Sprechergruppe-Information SGI und Erkennen des Kontexts unter Ausnutzung der Sprachinformation SI und Erzeugen der den erkannten Kontext repräsentierenden Kontext-Informationen CI und Erkennen der zu der Sprachinformation SI korrespondierenden Textinformation TI unter fortwährender Berücksichtigung der Segmentierung-Information ASI und der Sprache-Information LI und der Sprechergruppe-Information SGI und der Kontext-Information CI, wobei auf das Erzeugen der Informationen ASI, LI, SGI und CI und insbesondere auf das Berücksichtigen der dazu jeweils benötigten Informationen CHI, ASI, LI und SGI nachfolgend noch im Detail eingegangen ist.

Weiters wird bei dem Spracherkennungsverfahren die Sprachinformation SI empfangen und unter Ausnutzung des über einen der vier Empfangskanäle 3, 5, 6 oder 7 charakterisierenden Audiosignals AS der jeweils zum Empfangen der Sprachinformation SI verwendete Empfangskanal 3, 5, 6 oder 7 erkannt und eine den erkannten Empfangskanal 3, 5, 6 oder 7 repräsentierende Kanalangabe-Information CHI erzeugt und die Kanalangabe-Information CHI bei dem Erkennen der akustischen Segmentierung, der Sprache, der Sprecher-Gruppe, des Kontexts und der Textinformation TI berücksichtigt, wobei das Erkennen des Empfangskanals 3, 5, 6 oder 7 fortwährend, und zwar frameweise jeweils für den ersten Frame des jeweiligen Teils der Sprachinformation SI erfolgt, und korrespondierend dazu die Kanalangabe-Information CHI fortwährend aktualisiert, also neu erzeugt wird, und auch fortwährend berücksichtigt wird.

Bei dem Spracherkennungsverfahren erfolgt weiters das Erkennen der akustischen Segmentierung unter Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Kanalangabe-Information CHI. Dabei erfolgt das Erkennen der akustischen Segmentierung für den ersten Frame des jeweiligen Teils der Sprachinformation SI zeitverzögert um mindestens die zum Erzeugen der Kanalangabe-Information CHI benötigte Zeitspanne, während der der jeweilige Teil der Sprachinformation SI zum Erzeugen der Kanalangabe-Informationen CHI für den ersten Frame des jeweiligen Teils ausnutzbar ist. Eine weitere Verzögerung ist durch die von den ersten Spracheigenschaft-Erkennungsmitteln 20 verursachte zweite Verarbeitungsverzögerung D2 bedingt. Nachfolgend daran wird die akustische Segmentierung frameweise aktualisiert.

Bei dem Spracherkennungsverfahren erfolgt weiters das Erkennen der Sprache unter zusätzlicher Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Segmentierung-Information ASI. Dabei erfolgt das Erkennen der Sprache für den ersten Frame des jeweiligen Teils der Sprachinformation SI zeitverzögert um mindestens die zum Erzeugen der Kanalangabe-Information CHI und der Segmentierung-Information ASI benötigten Zeitspannen, während denen der jeweilige Teil der Sprachinformation SI zum Erzeugen der beiden Informationen CHI und ASI für den ersten Frame des jeweiligen Teils ausnutzbar ist. Eine weitere Verzögerung ist durch die von den zweiten Spracheigenschaft-Erkennungsmitteln 21 verursachte dritte Verarbeitungsverzögerung D3 bedingt. Nachfolgend daran wird die Sprache frameweise aktualisiert.

Bei dem Spracherkennungsverfahren erfolgt weiters das Erkennen der Sprecher-Gruppe unter zusätzlicher Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Segmentierung-Information ASI und Sprache-Information LI. Dabei erfolgt das Erkennen der Sprecher-Gruppe für den ersten Frame des jeweiligen Teils der Sprachinformation SI zeitverzögert um mindestens die zum Erzeugen der Kanalangabe-Information CHI, der Segmentierung-Information ASI und der Sprache-Information LI benötigten Zeitspannen, während denen der jeweilige Teil der Sprachinformation SI zum Erzeugen der Informationen CHI, ASI und LI für den ersten Frame des jeweiligen Teils ausnutzbar ist. Eine weitere Verzögerung ist durch die von den dritten Spracheigenschaft-Erkennungsmitteln 22 verursachte vierte

Verarbeitungsverzögerung D4 bedingt. Nachfolgend daran wird die Sprecher-Gruppe frameweise aktualisiert.

Bei dem Spracherkennungsverfahren erfolgt weiters das Erkennen des Kontexts unter zusätzlicher Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Segmentierung-Information ASI, Sprache-Information LI und Sprechergruppe-Information SGI. Dabei erfolgt das Erkennen des Kontexts für den ersten Frame des jeweiligen Teils der Sprachinformation SI zeitverzögert um mindestens die zum Erzeugen der Information CHI, ASI, LI und SGI benötigten Zeitspannen, während denen der jeweilige Teil der Sprachinformation SI zum Erzeugen der Informationen CHI, ASI, LI und SGI für den Teilbereich des jeweiligen Teils ausnutzbar ist. Eine weitere Verzögerung ist durch die von den vierten Spracheigenschaft-Erkennungsmitteln 23 verursachte fünfte Verarbeitungsverzögerung D5 bedingt. Nachfolgend daran wird der Kontext frameweise aktualisiert.

Bei dem Spracherkennungsverfahren erfolgt weiters das Erkennen der zu der Sprachinformation SI korrespondierenden Textinformation TI unter Berücksichtigung der zu jedem Frame des jeweiligen Teils der Sprachinformation SI korrespondierenden Information CHI, ASI, LI, SGI und CI für den ersten Frame des jeweiligen Teils der Sprachinformation SI zeitverzögert um mindestens die zum Erzeugen der Kanalangabe-Information CHI, der Segmentierung-Information ASI, der Sprache-Information LI, der Sprechergruppe-Information SGI und der Kontext-Information CI benötigten Zeitspannen, während denen der jeweilige Teil der Sprachinformation SI zum Erzeugen der Informationen CHI, ASI, LI, SGI und CI für den ersten Frame des jeweiligen Teils ausnutzbar ist. Eine weitere Verzögerung ist durch die von den Sprach-Erkennungsmitteln 24 verursachte sechste Verarbeitungsverzögerung D6 bedingt. Nachfolgend daran wird die Textinformation TI frameweise aktualisiert.

Mit dem Computer 1A wird das Spracherkennungsverfahren durchgeführt, wenn das Computerprogrammprodukt auf dem Computer 1A abgearbeitet wird. Das Computerprogrammprodukt ist auf einem computerlesbaren in der Figur 1 nicht dargestellten Medium gespeichert, welches Medium im vorliegenden Fall durch eine Compact-Disc (CD) realisiert ist. Es sei an dieser Stelle erwähnt, dass auch eine DVD oder ein bandartiger Datenträger oder eine Harddisc als Medium vorgesehen sein kann. Der Computer weist im vorliegenden Fall als die Recheneinheit einen einzigen Mikroprozessor

auf. Es sei jedoch erwähnt, dass aus Performance-Gründen auch mehrere Mikroprozessoren, beispielsweise für jedes Erkennungsmittel 18, 20, 21, 22, 23 und 24 ein eigener Mikroprozessor, vorgesehen sein können. Der interne Speicher 1B des Computers 1A ist im vorliegenden Fall durch eine Kombination einer in der Figur 1 nicht dargestellten Harddisc und eines mit Hilfe von sogenannten RAM-Speichern realisierten Arbeitsspeichers 39 realisiert, so dass das Computerprogramm-Produkt von dem computerlesbaren Medium zunächst auf die Harddisc speicherbar ist und zum Abarbeiten mit Hilfe der Recheneinheit in den Arbeitsspeicher 39 ladbar ist, wie dies dem Fachmann hinlänglich bekannt ist. Der Speicher 1B ist weiters zum Speichern des vorverarbeiteten Audiosignals PAS und der Informationen CHI, ASI, LI, SGI und CI und zum Speichern von in der Figur 1 nicht dargestellten zeitlichen Beziehungsdaten ausgebildet. Die zeitlichen Beziehungsdatenrepräsentieren eine zeitliche Beziehung zwischen den Teilbereichen der Sprachinformation SI, und den jeweils zu diesen Teilbereich korrespondierenden Informationen CHI, ASI, LI, SGI und CI, um das zeitlich synchronisierte Erkennen der akustischen Segmentierung, der Sprache, der Sprecher-Gruppe, des Kontexts bzw. der Textinformation TI für den jeweiligen Teilbereich der Sprachinformation SI zu ermöglichen.

Durch das Vorsehen der erfindungsgemäßen Maßnahmen ist auf vorteilhafte Weise erreicht, dass die Spracherkennungseinrichtung 1 bzw. das Spracherkennungsverfahren erstmals in einem Anwendungsfall einsetzbar ist, in dem gleichzeitig eine Vielzahl von die Sprachinformation SI charakterisierenden Spracheigenschaften einer im wesentlichen zu beliebigen Zeitpunkten auftretenden Veränderung unterworfen sind. Ein solcher Anwendungsfall ist beispielsweise bei einem Konferenz-Transkriptionssystem gegeben, bei dem eine von beliebigen Konferenzteilnehmern erzeugte Sprachinformation SI kontinuierlich und annähernd in Echtzeit in eine Textinformation TI umgewandelt werden muss, wobei die Konferenzteilnehmer die Sprachinformation SI in einem Konferenzraum mit Hilfe des Audiosignals AS über den ersten Empfangskanal 3 der Spracherkennungseinrichtung 1 zuführen. Dabei können die Konferenzteilnehmer verschiedene Sprachen verwenden und individuell unterschiedlichen Sprecher-Gruppen zugeordnet sein. Weiters können während einer Konferenz Umstände eintreten, wie beispielsweise Hintergrundgeräusche, welche die akustische Segmentierung beeinflussen. Weiters kann sich auch der jeweils verwendete

Kontext während der Konferenz verändern. Zusätzlich ist auf vorteilhafte Weise ermöglicht, dass auch Konferenzteilnehmer, die sich nicht im Konferenzraum befinden, über weitere Empfangskanäle 5, 6 oder 7 der Spracherkennungseinrichtung 1 die ihnen zugeordnete Sprachinformation SI zuführen können. Selbst in diesem Fall ist bei der Spracherkennungseinrichtung 1 ein zuverlässiges Erkennen der Textinformation TI gewährleistet, weil der jeweils verwendete Empfangskanal 3, 5, 6 oder 7 erkannt und bei dem Erkennen der Spracheigenschaften – also bei dem Erzeugen und Aktualisieren der Informationen CHI, ASI, LI, SGI und CI – bzw. bei dem Erkennen der Textinformation TI fortwährend berücksichtigt wird.

Weiters ist ein solcher Anwendungsfall dann gegeben, wenn beispielsweise bei einem sogenannten Callcenter Anrufe von beliebigen Personen, die sich unterschiedlicher Sprachen bedienen können, mitprotokolliert werden sollen.

Weiters ist ein solcher Anwendungsfall dann gegeben, wenn beispielsweise bei einem automatischen Telefon-Informationsdienst beliebige Anrufer bedient werden sollen. Es sei an dieser Stelle ausdrücklich darauf hingewiesen, dass die hier angeführten Anwendungsfälle keine vollständige Aufzählung darstellen.

Die in der Figur 3 dargestellten Featurevektor-Extrahierungsmittel 19 weisen eine Preemphasis-Stufe 40 auf, die zum Empfangen des Audiosignals AS und zum Abgeben eines das Audiosignal AS repräsentierenden modifizierten Audiosignals AS'' ausgebildet ist, wobei in dem modifizierten Audiosignal AS'' höhere Frequenzen betont sind, um den Frequenzgang zu nivellieren. Weiters ist eine Frameblocking-Stufe 41 vorgesehen, die zum Empfangen des modifizierten Audiosignals AS'' und zum Abgeben von in Frames F eingebetteten Teilen des modifizierten Audiosignals AS'' ausgebildet ist. Dabei weisen die benachbarten Frames F eine zeitliche Überlappung des Audiosignals AS'' in ihren Randbereichen auf. Weiters ist eine Windowing-Stufe 42 vorgesehen, die zum Empfangen der Frames F und zum Erzeugen von die Frames F repräsentierenden modifizierten Frames F', die hinsichtlich der Bandbreite des durch die Frames F repräsentierten Audiosignals begrenzt sind, um bei einer nachfolgenden Konversion in die Spektralebene unerwünschte Effekte zu vermeiden. Bei der Windowing-Stufe 42 kommt im vorliegenden Fall ein sogenanntes Hemming-Fenster zum Einsatz. Es sei jedoch erwähnt, dass auch andere Fenstertypen einsetzbar sind. Weiters ist eine Fast-Fourier–Transformation-Stufe 43 vorgesehen, die zum Empfangen der modifizierten Frames F' und

zum Erzeugen von zu dem in den modifizierten Frames F' enthaltenen bandbreitebegrenzten Audiosignals A'' korrespondierenden Vektoren V1 auf der Spektralebene ausgebildet ist, wobei im vorliegenden Fall ein sogenanntes „Zero-Pading"-Verfahren zum Einsatz kommt. Weiters ist eine Logarithmus-Filterbank-Stufe 44 vorgesehen, die zum Empfangen der ersten Vektoren V1 und der Kanalangabe-Information CHI und unter Ausnutzung der ersten Vektoren V1 und unter Berücksichtigung der Kanalangabe-Information CHI zum Erzeugen und Abgeben von zweiten Vektoren V2 ausgebildet ist, wobei die zweiten Vektoren V2 eine logarithmische Abbildung von aus den ersten Vektoren V1 mit Hilfe einer Filterbankmethode erzeugbaren Zwischenvektoren repräsentieren.

Die in der Figur 12 dargestellte Logarithmus-Filterbank-Stufe 44 weist eine Filterbankparameterpool-Stufe 44A auf, die einen Pool von Filterbankparametern speichert. Weiters ist eine Filterparameter-Auswahlstufe 44B vorgesehen, die zum Empfangen der Kanalangabe-Information CHI und zum Auswählen von zu der Kanalangabe-Information CHI korrespondierenden Filterbankparametern FP ausgebildet ist. Weiters ist ein sogenannter Logarithmus-Filterbank-Kern 44C vorgesehen, der in Abhängigkeit von den von der Filterparameter-Auswahlstufe 44B empfangbaren Filterbankparametern FP zum Verarbeiten der ersten Vektoren V1 und zum Erzeugen der zweiten Vektoren V2 ausgebildet ist.

Die in der Figur 3 dargestellten Featurevektor-Extrahierungsmittel 19 weisen weiters eine erste Normierung-Stufe 45 auf, die zum Empfangen der zweiten Vektoren V2 und zum Erzeugen und Abgeben von hinsichtlich der Amplitude der zweiten Vektoren V2 mittelwertfreien dritten Vektoren V3 ausgebildet sind. Dadurch ist gewährleistet, dass eine vom jeweiligen Empfangskanal unabhängige Weiterverarbeitung ermöglicht ist. Weiters ist eine zweite Normierung-Stufe 46 vorgesehen, die zum Empfangen der dritten Vektoren V3 und unter Berücksichtigung der zeitlichen Varianz für jede der Komponenten der dritten Vektoren V3 zum Erzeugen von hinsichtlich der zeitlichen Varianz der dritten Vektoren V3 normierten vierten Vektoren V4 ausgebildet ist. Weiters ist eine Diskrete-Cosinus-Transformation-Stufe 47 vorgesehen, die zum Empfangen der vierten Vektoren V4 und zum Umwandeln der vierten Vektoren V4 in die sogenannte „Cepstral"-Ebene und zum Abgeben von fünften Vektoren V5 ausgebildet sind, die zu den vierten Vektoren V4 korrespondieren. Weiters ist eine Feature-Vektor-Erzeugungsstufe 48 vorgesehen, die zum

Empfangen der fünften Vektoren V5 und zum Erzeugen der ersten und der zweiten zeitlichen Ableitung der fünften Vektoren V5 ausgebildet sind, so dass die von der Featurevektor-Erzeugungsstufe 48 abgebbare vektormäßige Repräsentation des Audiosignal AS in Form der Featurevektoren FV, die fünften Vektoren V5 in der „Cepstral"-Ebene und die dazu korrespondierenden zeitlichen Ableitungen aufweist.

Die in der Figur 4 dargestellten Empfangskanal-Erkennungsmittel 18 weisen eingangsseitig eine Spektralvektor-Extrahierungsstufe 49 auf, die zum Empfangen des Audiosignals AS und zum Extrahieren und Abgeben von Spektralvektoren V6 ausgebildet ist, welche Spektralvektoren V6 das Audiosignal AS auf der Spektralebene repräsentieren. Weiters weisen die Empfangskanal-Erkennungsmittel 18 eine Bandbegrenzungs-Erkennungsstufe 50 auf, die zum Empfangen der Spektralvektoren V6 und unter Ausnutzung des Spektralvektoren V6 zum Erkennen einer Bandbegrenzung des Frequenzbandes des Audiosignals AS ausgebildet ist, wobei die jeweils festgestellte Bandbegrenzung für jeweils einen der vier Empfangskanäle repräsentativ ist. Die Bandbegrenzung-Erkennungsstufe 50 ist weiters zum Abgeben einer die erkannte Bandbegrenzung repräsentierenden Bandbegrenzung-Information BWI ausgebildet. Die Empfangskanal-Erkennungsmittel 18 weisen weiters eine Kanal-Klassifikationsstufe 51 auf, die zum Empfangen der Bandbegrenzung-Information BWI und unter Ausnutzung dieser Information BWI zum Klassifizieren des jeweils vorliegenden Empfangskanals und zum Erzeugen der dazu korrespondierenden Kanalangabe-Information CHI ausgebildet ist.

Die in der Figur 5 dargestellten ersten Spracheigenschaft-Erkennungsmittel 20 weisen eine Sprachpause-Erkennungsstufe 52 und eine Nicht-Sprache-Erkennungsstufe 53 und eine Musik-Erkennungsstufe 54 auf, wobei jeder der Erkennungsstufen 52, 53 und 54 die Featurevektoren FV zuführbar sind. Die Sprachpause-Erkennungsstufe 52 ist zum Erkennen von Sprachpausen repräsentierenden Featurevektoren FV und zum Abgeben einer das Erkennungsergebnis repräsentierenden Sprachpause-Information SI ausgebildet. Die Nicht-Sprache-Erkennungsstufe 53 ist zum Empfangen der Kanalangabeinformation CHI und unter Berücksichtigung der Kanalangabe-Information CHI zum Erkennen von Nicht-Sprache repräsentierenden Featurevektoren FV und zum Abgeben einer die Nicht-Sprache repräsentierenden Nicht-Sprache-Information NSI ausgebildet. Die Musik-Erkennungsstufe 54 ist zum Empfangen der Kanalangabeinformation CHI und unter Berücksichtigung der Kanalangabe-Information CHI zum Erkennen von Musik

repräsentierenden Featurevektoren FV und zum Erzeugen und Abgeben einer das Erkennen der Musik repräsentierenden Musik-Information MI ausgebildet. Die ersten Spracheigenschaft-Erkennungsmittel 20 weisen weiters eine Information-Auswertungsstufe 55 auf, die zum Empfangen der Sprachpause-Information SI und der Nicht-Sprache-Information NSI und der Musik-Information MI ausgebildet ist. Die Information-Auswertungsstufe 55 ist weiters zum Auswerten der Informationen SI, NSI und MI und als ein Ergebnis des Auswertens zum Erzeugen und zum Abgeben der Segmentierung-Information ASI ausgebildet, wobei die Segmentierung-Information ASI angibt, ob der jeweils durch die Featurevektoren FV repräsentierte Frame des Audiosignals AS einer Sprachpause oder Nicht-Sprache oder Musik zugeordnet ist, und die angibt, wenn der jeweilige Frame weder einer Sprachpause oder einer Nicht-Sprache oder einer Musik zugeordnet ist, dass der jeweilige Frame Sprache zugeordnet ist.

Die in der Figur 13 im Detail dargestellte Musik-Erkennungsstufe 54 ist weiters auf trainierbare Weise zum Erkennen von Musik ausgebildet und ist zu diesem Zweck zum Empfangen einer Segmentierung-Training-Information STI ausgebildet. Die Musik-Erkennungsstufe 54 weist eine Klassifikationsstufe 56 auf, die unter Zuhilfenahme von zwei Gruppen von sogenannten „Gausian-Mixture-Modells" zum Klassifizieren der Featurevektoren FV hinsichtlich von Musik repräsentierenden Featurevektoren FV und hinsichtlich von Nicht-Musik repräsentierenden Featurevektoren FV ausgebildet sind. Dabei ist jedes zu der ersten Gruppe gehörende erste Gausian-Mixture-Modell GMM1 einer Musikklassifizierung und jedes zu der zweiten Gruppe gehörende zweite Gausian-Mixture-Modell GMM2 einer Nichtmusikklassifizierung zugeordnet. Die Klassifikationsstufe 56 ist weiters als ein Ergebnis des Klassifizierens zum Abgeben der Musikinformation MI ausgebildet ist. Die Musik-Erkennungsstufe 54 weist weiters eine erste Modell-Auswahlstufe 57 und eine erste Modell-Speicherstufe 58 auf. Die erste Modell-Speicherstufe 58 ist für jeden der Empfangkanäle zum Speichern eines der Musikklassifizierung zugeordneten Gausian-Mixture-Modells GMM1 und zum Speichern eines der Nichtmusikklassifizierung zugeordneten Gausian-Mixture-Modell GMM2 ausgebildet. Die erste Modell-Auswahlstufe 57 ist zum Empfangen der Kanalangabe-Information CHI und unter Zuhilfenahme der Kanalangabe-Information CHI zum Auswählen eines zu dem jeweils angegebenen Empfangskanal korrespondierenden Paares von Gausian-Mixture-Modells GMM1 und GMM2 und zum Abgeben der kanalspezifisch

ausgewählten Gausian-Mixture-Modells GMM1 und GMM2 an die Klassifikationsstufe 56 ausgebildet.

Die Musik-Erkennungsstufe 54 ist weiters zum Trainieren der Gausian-Mixture-Modells ausgebildet und weist zu diesem Zweck eine erste Trainingsstufe 59 und eine erste Datenstromsteuerstufe 60 auf. Der ersten Trainingsstufe 59 sind bei dem Training mit Hilfe der Datenstromsteuerstufe 60 Featurevektoren FV zuführbar, die in vorbestimmter Weise jeweils zu einer einzigen Klasse, nämlich Musik oder Nicht-Musik gehören. Die Trainingsstufe 59 ist zum Trainieren der kanalspezifischen Paare von Gausian-Mixture-Modells GMM1 und GMM2 ausgebildet. Die erste Modell-Auswahlstufe 57 ist unter Zuhilfenahme der Kanalangabe-Information CHI und der Segmentierung-Training-Information STI zum Abgeben der Gausian-Mixture-Modells GMM1 und GMM2 an dafür vorgesehene Speicherpositionen in der ersten Modell-Speicherstufe 58 ausgebildet.

Die in der Figur 6 dargestellten zweiten Spracheigenschaft-Erkennungsmittel 21 weisen eingangsseitig eine erste Sprachfilterstufe 61 auf, die zum Empfangen der Featurevektoren FV und zum Empfangen der Segmentierung-Information ASI und unter Ausnutzung der Segmentierung-Information ASI zum Ausfiltern von Sprache repräsentierenden Featurevektoren FV und zum Abgeben der die Sprache repräsentierenden Featurevektoren FV ausgebildet ist. Die zweiten Spracheigenschaft-Erkennungsmittel 21 weisen weiters eine zweite Modell-Speicherstufe 62 auf, die für jeden der vier Empfangskanäle zum Speichern von jeweils einem mehrsprachigen ersten Phonem-Modell PM1 ausgebildet und vorgesehen ist. Die Erkennungsmittel 21 weisen weiters eine zweite Modell-Auswahlstufe 63 auf, die zum Empfangen der Kanalangabe-Information CHI und unter Ausnutzung der Kanalangabe-Information CHI zum Zugreifen auf das zu dem durch die Kanalangabe-Information CHI angegebenen Empfangskanal korrespondierende mehrsprachige Phonem-Modell PM1 in den zweiten Modell-Speicherstufe 62 und zum Abgeben des so ausgewählten kanalspezifischen mehrsprachigen Phonem-Modells PM1 ausgebildet ist. Die Erkennungsmittel 21 weisen weiters eine Phonem-Erkennungsstufe 64 auf, die zum Empfangen der Sprache repräsentierenden Featurevektoren FV und des Phonem-Modells PM1 und unter Ausnutzung der Featurevektoren FV und des Phonem-Modells PM1 zum Erzeugen und zum Abgeben einer phonetischen Transkription PT der durch die Featurevektoren FV repräsentierten Sprache

ausgebildet ist. Die Erkennungsmittel 21 weisen weiters eine dritte Modell-Speicherstufe 65 auf, die für jede Sprache zum Speichern eines phonotaktischen Modells PTM ausgebildet und vorgesehen ist. Die Erkennungsmittel 21 weisen weiters eine zweite Klassifikationsstufe 66 auf, die zum Zugreifen auf die dritte Modell-Speicherstufe 65 und unter Zuhilfenahme des phonotaktischen Modells PTM zum phonotaktischen Klassifizieren der phonetischen Transkription PT ausgebildet sind, wobei die Wahrscheinlichkeit des Vorliegens einer Sprache für jede verfügbare Sprache bestimmbar ist. Als ein Ergebnis des Bestimmens der zu jeder Sprache korrespondierenden Wahrscheinlichkeiten ist die zweite Klassifikationsstufe 66 zum Erzeugen und zum Abgeben der Sprache-Information LI ausgebildet, welche Sprache-Information LI jene Sprache angibt, für die die größte Wahrscheinlichkeit festgestellt wurde.

Die Erkennungsmittel 21 sind weiters auf trainierbare Weise hinsichtlich des Erkennens der Sprache beeinflussbar und weisen zu diesem Zweck eine zweite Datenstromsteuerstufe 67, eine dritte Datenstromsteuerstufe 68 und eine zweite Trainingsstufe 69 und eine dritte Trainingsstufe 70 auf. Im Falle eines Trainings sind mit Hilfe der zweiten Datenstromsteuerstufe 67 die Sprache repräsentierenden Featurevektoren FV der zweiten Trainingsstufe 69 zuführbar. Die zweite Trainingsstufe 69 ist zum Empfangen dieser Featurevektoren FV und zum Empfangen einer Training-Text-Information TTI und zum Empfangen der Kanalangabe-Information CHI ausgebildet, wobei eine aus der Training-Text-Information TTI erzeugte phonetischen Transkription zu der durch die Featurevektoren FV repräsentierten Sprache korrespondiert. Die zweite Trainingsstufe 69 ist demgemäß unter Ausnutzung der Featurevektoren FV und der Training-Text-Information TTI zum Trainieren und zum Abgeben des trainierten Phonem-Modells PM1 an die Modell-Auswahlstufe 63 ausgebildet. Die Modell-Auswahlstufe 63 ist weiters unter Zuhilfenahme der Kanalangabe-Information CHI zum Abgeben des trainierten Phonem-Modells PM1 an die zweite Modell-Speicherstufe 62 ausgebildet, wo sie an einer zu der Kanalangabe-Information CHI korrespondierenden Speicherposition in der zweiten Modell-Speicherstufe 62 speicherbar sind.

Im Fall des Trainings ist weiters mit Hilfe der dritten Datenstromsteuerstufe 68 die von der Phonem-Erkennungsstufe 64 erzeugbare phonetische Transkription PT der dritten Trainingsstufe 70 zuführbar. Die dritte Trainingsstufe 70 ist zum Empfangen der phonetischen Transkription PT und zum Trainieren und zum Abgeben eines zu der

jeweiligen Training-Sprache-Information TLI zugeordneten phontaktischen Modells PTM an die dritte Modell-Speicherstufe 65 ausgebildet. Die dritte Modell-Speicherstufe 65 ist zum Speichern des zu einer Sprache gehörenden phonotaktischen Modells PTM an einer zu der Training-Sprache-Information TLI korrespondierenden Speicherposition ausgebildet. Es sei an dieser Stelle erwähnt, dass die in der zweiten Modell-Speicherstufe 62 und in der dritten Modell-Speicherstufe 65 gespeicherten Modelle PM1 und PTM im Fachjargon als trainierbare Ressourcen bezeichnet werden.

Die zweite Trainingsstufe 69 ist im Detail in der Figur 14 dargestellt und weist eine vierte Modell-Speicherstufe 71 und eine dritte Modell-Auswahlstufe 72 und eine Modell-Gruppierungsstufe 73 und eine Modell-Ausrichtungsstufe 74 und eine Modell-Abschätzungsstufe 75 auf. Die vierte Modell-Speicherstufe 71 ist für jeden Kanal und jede Sprache zum Speichern eines kanal- und sprachespezifischen initialen Phonem-Modells IPM ausgebildet und vorgesehen. Die dritte Modell-Auswahlstufe 72 ist zum Zugreifen auf die vierte Modell-Speicherstufe 71 und zum Empfangen der Kanalangabe-Information CHI und unter Ausnutzung der Kanalangabe-Information CHI zum Auslesen des zu der Kanalangabe-Information CHI korrespondierenden initialen Phonem-Modells IPM für alle Sprachen ausgebildet. Die dritte Modell-Auswahlstufe 72 ist weiters zum Abgeben von einer zu dem jeweiligen Kanal korrespondierenden Mehrzahl von sprachespezifischen Phonem-Modellen IPM an die Modell-Gruppierungsstufe 73 ausgebildet. Die Modell-Gruppierungsstufe 73 ist zum Gruppieren von zueinander ähnlichen und zu verschiedenen Sprachen gehörenden sprachspezifischen Phonem-Modellen IPM und zum Erzeugen und zum Abgeben eines initialen mehrsprachigen Phonem-Modells IMPM an die Modell-Ausrichtungsstufe 74 ausgebildet. Die Modell-Ausrichtungsstufe 74 ist zum Empfangen der Sprache repräsentierenden Featurevektoren FV und zum Empfangen der dazu korrespondierenden Training-Text-Information TTI und unter Zuhilfenahme des initialen mehrsprachigen Phonem-Modells IMPM zum Erzeugen von Zuordnungsinformationen RE ausgebildet, die zum Zuordnen der Featurevektoren FV zu durch die Training-Text-Information TTI repräsentierten Textteilen vorgesehen sind, wobei die Zuordnungsinformationen RE im Fachjargon auch als Pfade bezeichnet werden. Das Zuordnen selbst wird im Fachjargon auch als „alignment" bezeichnet. Von der Modell-Ausrichtungsstufe 74 sind die Zuordnungsinformationen RE und die Featurevektoren FV an die Modell-Abschätzungsstufe 75 abgebbar. Die Modell-Abschätzungsstufe 75 ist unter

Ausnutzung der Zuordnungsinformationen RE und der Featurevektoren FV zum Erzeugen und Abgeben des auf dem initialen mehrsprachigen Phonem-Modell IMPM basierenden mehrsprachigen Phonem-Modell PM1 an die in der Figur 7 dargestellte zweite Modell-Speicherstufe 62 ausgebildet. Zu diesem Zweck wird unter Ausnutzung der Featurevektoren FV und der Zuordnungsinformation RE ein temporäres mehrsprachiges Phonem-Modell TMPM erzeugt und an die Modell-Abschätzungsstufe 74 abgegeben, wobei in mehreren Iterationsschritten, also durch ein mehrfaches Zusammenwirken der Stufen 74 und 75, das mehrsprachige Phonem-Modell PM1 erzeugt wird.

Die in der Figur 7 im Detail dargestellten dritten Spracheigenschaft-Erkennungsmittel 22 weisen eingangsseitig eine zweite Sprachfilterstufe 76 auf, die zum Empfangen der Featurevektoren FV und der Segmentierung-Information ASI und unter Ausnutzung der Segmentierung-Information ASI zum Filtern und Abgeben von Sprache repräsentierenden Featurevektoren FV ausgebildet ist. Die Erkennungsmittel 22 weisen weiters eine fünfte Modell-Speicherstufe 77 auf, die für jeden Kanal und jede Sprache zum Speichern von Sprechergruppen-Modellen SGM ausgebildet und vorgesehen ist. Die Erkennungsmittel 22 weisen weiters eine vierte Modell-Auswahlstufe 78 auf, die zum Empfangen der Kanalangabe-Information CHI und der Sprache-Information LI ausgebildet ist und die unter Ausnutzung der Kanalangabe-Information CHI und der Sprache-Information LI zum Zugreifen auf das jeweilige Sprechergruppen-Modell SGM, das zu der jeweiligen Kanalangabe-Information CHI und der jeweiligen Sprache-Information LI korrespondiert, ausgebildet ist. Die vierte Modell-Auswahlstufe 78 ist weiters zum Abgeben des durch das Zugreifen auf die fünfte Modell-Speicherstufe 77 auslesbaren Sprechergruppe-Modells SGM ausgebildet. Die Erkennungsmittel 22 weisen weiters eine dritte Klassifikationsstufe 79 auf, die zum Empfangen des von der vierten Modell-Auswahlstufe 78 in Abhängigkeit von den Informationen CHI und LI ausgewählten Sprechergruppe-Modells SGM und zum Empfangen der Sprache repräsentierenden Featurevektoren FV und unter Zuhilfenahme des ausgewählten Sprechergruppen-Modells SGM zum Klassifizieren, welcher Sprechergruppe die Featurevektoren FV zuordenbar sind, ausgebildet ist. Die dritte Klassifikationsstufe 79 ist weiters als ein Ergebnis des Klassifizierens zum Erzeugen und zum Abgeben der Sprechergruppe-Information SGI ausgebildet.

Mit Hilfe der fünften Modell-Speicherstufe 77 ist eine weitere trainierbare

Ressource realisiert, wobei die darin gespeicherten Sprechergruppen-Modelle SGM auf trainierbare Weise veränderbar sind. Zu diesem Zweck weisen die Erkennungsmittel 22 eine vierte Trainingstufe 80 und eine vierte Datenstromsteuerstufe 81 auf. Im Fall eines Trainings sind mit Hilfe der vierten Datenstromsteuerstufe 81 die Sprache repräsentierende Featurevektoren FV der vierten Trainingsstufe 80 zuführbar. Die vierte Trainingsstufe 80 ist für eine Anzahl von Sprecher zum Empfangen von jeweils einem Sprecher zugeordneten Featurevektoren FV und der jeweils dazu korrespondierenden Training-Text-Information TTI und zum Trainieren des jeweiligen Sprechergruppen-Modells SGM und zum Abgeben des jeweiligen trainierten Sprechergruppen-Modells SGM an die vierte Modell-Auswahlstufe 78 ausgebildet.

Die in der Figur 15 im Detail dargestellte vierte Training-Stufe 80 weist eine sechste Modell-Speicherstufe 82, eine fünfte Modell-Auswahlstufe 83, eine Modell-Anpassungsstufe 84, eine Zwischenspeicherstufe 85 und eine Modell-Gruppierungsstufe 86 auf. Die sechste Modell-Speicherstufe 82 ist für jeden Kanal und jede Sprache zum Speichern von sprecherunabhängigen Phonem-Modellen SIPM vorgesehen und ausgebildet. Die fünfte Modell-Auswahlstufe 83 ist zum Empfangen der Kanalangabe-Information CHI und der Sprache-Information LI und unter Ausnutzung dieser beiden Informationen CHI und LI zum Zugreifen auf die sechste Modell-Speicherstufe 82 bzw. auf das zu der jeweiligen Information CHI und LI korrespondierende initiale sprecherunabhängige Phonem-Modell SIPM und zum Abgeben des ausgewählten nunmehr kanal- und sprachespezifischen und sprecherunabhängigen Phonem-Modells SIPM ausgebildet.

Die Modell-Anpassungsstufe 84 ist zum Empfangen des gemäß der Kanalangabe-Information CHI und der Sprache-Information LI ausgewählten und somit kanal- und sprachespezifischen initialen sprecherunabhängigen Phonem-Modells SIPM, der Sprache repräsentierenden Featurevektoren FV und der dazu korrespondierenden Training-Text-Information TTI ausgebildet. Die Modell-Anpassungsstufe 84 ist weiters für eine Vielzahl von Sprechern, deren Sprachinformation SI durch die Featurevektoren FV repräsentiert ist, zum Erzeugen und zum Abgeben von je einem Sprechermodell SM an die Zwischenspeicherstufe 85 ausgebildet, bei der das jeweilige Sprechermodell SM speicherbar ist. Das Sprachmodell SM wird auf Grundlage des sprecheruzabhängigen Phonem-Modells SIPM unter Anwendung eines Adaptionsverfahrens erzeugt. Nachdem

für die gesamte Anzahl der Sprecher die Sprechermodelle SM gespeichert wurden, ist mit Hilfe der Modell-Gruppierungsstufe 86 ein Gruppieren der Vielzahl der Sprechermodelle SM hinsichtlich ähnlicher Sprechereigenschaften zu einzelnen Sprechergruppen-Modellen SGM durchführbar. Die einzelnen Sprechergruppen-Modelle SGM sind an die Modell-Auswahlstufe 78 abgebbar und von der Modell-Auswahlstufe 78 unter Ausnutzung der Informationen CHI und LI in der Modell-Speicherstufe 77 speicherbar.

Die in der Figur 8 im Detail dargestellten vierten Spracheigenschaft-Erkennungsmittel 23 weisen eine Stichwort-Phonem-Sequenzerkennungsstufe 88 und eine Stichwort-Erkennungsstufe 89 und eine Stichwort-Kontext-Zuordnungsstufe 90 auf. Die Stufe 88 ist zum Empfangen der Featurevektoren FV und zum Empfangen eines zweiten Phonem-Modells PM2, das kanal- und sprache- und sprechergruppespezifisch ist, und zum Empfangen einer Stichwort-Lexikon-Information KLI ausgebildet. Die Stufe 88 ist weiters unter Ausnutzung des zweiten Phonem-Modells PM2 und der Stichwort-Lexikon-Information KLI zum Erkennen einer durch die Featurevektoren FV repräsentierten Stichwort-Sequenz und zum Erzeugen und zum Abgeben einer Stichwort-Bewertung-Information KSI ausgebildet, die ein erkanntes Stichwort und die Wahrscheinlichkeit, mit der dieses Stichwort erkannt wurde, repräsentiert. Die Stichwort-Erkennungsstufe 89 ist zum Empfangen der Stichwort-Bewertung-Information KSI und zum Empfangen eines von dem Empfangskanal, der Sprache, der Sprechergruppe und dem Stichwort abhängigen Stichwort-Entscheidung-Schwellwerts KWDT ausgebildet. Die Stufe 89 ist weiters unter Zuhilfenahme des Stichwort-Entscheidung-Schwellwerts KWDT zum Erkennen ausgebildet, welche der mit Hilfe der Stichwort-Bewertungsinformation KSI empfangenen Stichwörter erkannt wurden. Als ein Ergebnis dieses Erkennens ist die Stichwort-Erkennungsstufe 89 zum Erzeugen einer Stichwort-Information KWI und zum Abgeben dieser Stichwort-Information KWI an die Stichwort-Kontext-Zuordnungsstufe 90 ausgebildet. Die Stichwort-Kontext-Zuordnungsstufe 90 ist weiters zum Zuordnen des mit Hilfe der Stichwort-Information KWI empfangenen Stichwortes zu einem Kontext ausgebildet, der im Fachjargon oft auch als „topic" bezeichnet wird. Als ein Ergebnis dieses Zuordnens ist die Stichwort-Kontext-Zuordnungsstufe 90 zum Erzeugen der Kontext-Information CI ausgebildet. Die vierten Spracheigenschaft-Erkennungsmittel 23 weisen weiters eine siebente Modell-Speicherstufe 91 auf, die für jeden Empfangskanal und jede Sprache und jede Sprecher-Gruppe zum Speichern der zweiten Phonem-Modelle

PM2 ausgebildet und vorgesehen ist. Die Erkennungsmittel 23 weisen weiters eine sechste Modell-Auswahlstufe 92 auf, die zum Empfangen der Kanalangabe-Information CHI und der Sprache-Information LI und der Sprechergruppe-Information SGI ausgebildet ist. Die sechste Modell-Auswahlstufe 92 ist weiters unter Zuhilfenahme der Kanalangabe-Information CHI und der Sprach-Information LI und der Sprecher-Gruppe-Information SGI zum Auswählen von einem der in der siebenten Modell-Speicherstufe 91 gespeicherten zweiten Phonem-Modelle PM2 und zum Abgeben des ausgewählten zweiten Phonem-Modells PM2 an die Stichwort-Phonem-Sequenzerkennungsstufe 88 ausgebildet.

Die Erkennungsmittel 23 weisen weiters eine Schlüsselwort-Lexikon-Speicherstufe 93 und eine Sprache-Auswahlstufe 94 auf. Die Schlüsselwort-Lexikon-Speicherstufe 93 ist zu jeder verfügbaren Sprache zum Speichern von Schlüsselwörtern ausgebildet und vorgesehen. Die Sprache-Auswahlstufe 94 ist zum Empfangen der Sprache-Information LI und zum Zugreifen auf die Schlüsselwort-Lexikon-Speicherstufe 93 ausgebildet, wobei unter Zuhilfenahme der Sprache-Information LI eine zu der Sprache-Information LI korrespondierende Stichwort-Lexikon-Information KLI, welche die Stichwörter einer Sprache repräsentiert, an die Stichwort-Phonem-Sequenzerkennungsstufe 88 abgebbar ist. Die Erkennungsmittel 23 weisen weiters eine Schwellwert-Speicherstufe 95 auf, die zum Speichern von dem jeweiligen Empfangskanal, der Sprache, der Sprecher-Gruppe und dem Stichwort abhängigen Stichwort-Entscheidungs-Schwellwerten KWDT ausgebildet und vorgesehen ist. Die Erkennungsmittel 23 weisen weiters eine Schwellwert-Auswahlstufe 96 auf, die zum Empfangen der Kanalangabe-Information CHI und der Sprache-Information LI und der Sprechergruppe-Information SGI ausgebildet ist. Die Schwellwert-Auswahlstufe 96 ist weiters in Abhängigkeit von den Informationen CHI, LI und SGI zum Zugreifen auf die in der Schwellwert-Speicherstufe 95 gespeicherte zu den Informationen CHI, LI und SGI korrespondierenden Stichwort-Entscheidung-Schwellwerte KWDT ausgebildet. Die Schwellwert-Auswahlstufe 96 ist weiters zum Abgeben der so ausgewählten Stichwort-Entscheidung-Schwellwerte KWDT an die Stichwort-Erkennungsstufe 89 ausgebildet.

Die Erkennungsmittel 23 sind weiters auf trainierbare Weise zum Erkennen der Kontext-Information CI ausgebildet, wobei zwei trainierbare Ressourcen durch die siebente Modell-Speicherstufe 91 und die Schwellwert-Speicherstufe 95 gebildet sind. Die Erkennungsmittel 23 weisen weiters eine fünfte Trainingsstufe 97 und eine sechste

Trainingsstufe 98 und eine fünfte Datenstromsteuerstufe 99 und eine sechste Datenstromsteuerstufe 100 auf. Bei einem Training der Erkennungsmittel 23 sind mit Hilfe der sechsten Datenstromsteuerstufe 100 die Featurevektoren FV der fünften Trainingsstufe 97 zuführbar. Die fünfte Trainingsstufe 97 ist weiters zum Empfangen der Featurevektoren FV und der dazu korrespondierenden Training-Text-Information TTI und unter Zuhilfenahme eines sogenannte Viterbi-Algorithmus zum Erzeugen und zum Abgeben eines der zweiten Phonem-Modelle PM2 an die sechste Modell-Auswahlstufe 92 ausgebildet, wodurch die zweiten Phonem-Modelle PM2 für jeden Kanal und jede Sprache und jede Sprecher-Gruppe erzeugt werden. Mit Hilfe der Modell-Auswahlstufe 92 sind die zweiten Phonem-Modelle PM2 in der Modell-Speicherstufe 91 an mit Hilfe der Informationen CHI, LI und SGI bestimmbaren Speicherpositionen speicherbar. Weiters ist mit Hilfe der fünften Datenstromsteuerstufe 99 die Stichwort-Lexikon-Information KLI der sechsten Trainingsstufe 98 zuführbar. Bei einem Training ist die Stichwort-Phonem-Sequenzerkennungsstufe 88 zum Erkennen einer Phonem-Sequenz in Featurevektoren FV, welche die Sprache repräsentieren, und zum Erzeugen und zum Abgeben einer die erkannte Phonem-Sequenz repräsentierenden Phonem-Bewertung-Information PSI an die sechste Trainingsstufe 98 ausgebildet, wobei die Phonem-Bewertung-Information PSI die erkannten Phoneme und zu jedem Phonem die Wahrscheinlichkeit, mit der es erkannt wurde, repräsentiert.

Die sechste Trainingsstufe 98 ist zum Empfangen der Phonem-Bewertungs-Information PSI und der Stichwort-Lexikon-Information KLI und unter Ausnutzung dieser beiden Informationen PSI und KLI zum Erzeugen – also zum Trainieren - und zum Abgeben eines zu den Informationen CHI, LI und SGI korrespondierenden Stichwort-Entscheidung-Schwellwerts KWDT an die Schwellwert-Auswahlstufe 96 ausgebildet. Die Schwellwert-Auswahlstufe 96 ist unter Ausnutzung der Informationen CHI, LI und SGI zum Abgeben des Stichwort-Entscheidung-Schwellwerts KWDT an die Schwellwert-Speichermittel 95 ausgebildet. Mit Hilfe der Schwellwert-Auswahlstufe 96 ist der Stichwort-Entscheidung-Schwellwert KWDT in einer durch die Informationen CHI, LI und SGI bestimmten Speicherposition speicherbar.

Die in der Figur 16 im Detail dargestellte sechste Trainingsstufe 98 weist eine Phonem-Wahrscheinlichkeitsverteilung-Abschätzungsstufe 101 auf, die zum Empfangen der Phonem-Bewertung-Information PSI und zum Abschätzen einer statistischen

Verteilung der gesprochenen Phoneme und der nicht gesprochenen Phoneme unter der Annahme, dass es sich jeweils um eine Gauß-Verteilung handelt, ausgebildet ist. Die Stufe 101 ist also ein Ergebnis dieser Abschätzung zum Erzeugen und zum Abgeben einer ersten Abschätzung-Information E1 ausgebildet. Die sechste Trainingsstufe 98 weist weiters eine Stichwort-Wahrscheinlichkeitsverteilung-Abschätzungsstufe 102 auf, die zum Empfangen der ersten Abschätzung-Information E1 und der Stichwort-Lexikon-Information KLI ausgebildet ist. Die Stufe 102 ist weiters unter Ausnutzung der beiden Informationen KLI und EI zum Abschätzen einer statistischen Verteilung der gesprochenen Stichwörter und der nicht gesprochenen Stichwörter ausgebildet. Die Stufe 102 ist weiters als ein Ergebnis des Abschätzens zum Erzeugen und zum Abgeben einer zweiten Abschätzung-Information E2 ausgebildet. Die sechste Trainingsstufe 98 weist weiters eine Stichwort-Entscheidung-Schwellwert-Abschätzungsstufe 103 auf, die unter Ausnutzung der zweiten Abschätzung-Information E2 zum Abschätzen des jeweiligen Stichwort-Entscheidung-Schwellwerts KWDT und als ein Ergebnis dieses Abschätzens zum Abgeben des Stichwort-Entscheidung-Schwellwerts KWDT ausgebildet ist.

Die in der Figur 9 im Detail dargestellten Spracherkennungsmittel 24 weisen eingangsseitig eine dritte Sprachfilterstufe 104 auf, die zum Empfangen der Featurevektoren FV und zum Empfangen der Segmentierung-Information ASI und unter Ausnutzung der Segmentierung-Information ASI zum Filtern der empfangenen Featurevektoren FV und zum Abgeben von Sprache repräsentierenden Featurevektoren FV ausgebildet ist.

Die Erkennungsmittel 24 weisen weiters eine Sprachmuster-Erkennungsstufe 105 auf, die zum Empfangen der Sprache repräsentierenden Featurevektoren FV und zum Empfangen eines dritten Phonem-Modells PM3 und zum Empfangen von Kontext-Daten CD ausgebildet ist. Die Sprachmuster-Erkennungsstufe 105 ist weiters unter Ausnutzung des dritten Phonem-Modells PM3 und der Kontext-Daten CD zum Erkennen eines Musters in den Featurevektoren FV, welche die Sprache repräsentieren, und als ein Ergebnis des Erkennens eines solchen Musters zum Erzeugen und zum Abgeben einer Wortgraph-Information WGI ausgebildet. Die Wortgraph-Information WGI repräsentiert Graphen von Wörtern oder Wortfolgen und ihnen zugehörige Wahrscheinlichkeitsinformationen, die angeben, mit welcher Wahrscheinlichkeit die Wörter oder Wortfolgen in der jeweiligen gesprochenen Sprache möglicherweise auftreten.

Die Erkennungsmittel 24 weisen weiters eine Graph-Bewertungsstufe 106 auf, die zum Empfangen der Wortgraph-Information WGI und zum Feststellen ausgebildet ist, welcher Pfad in dem Graph die hinsichtlich des Erkennens der Textinformation TI die beste Wortfolge aufweist. Die Graph-Bewertungsstufe 106 ist weiters als ein Ergebnis des Feststellens der besten Wortfolge zum Abgeben einer zu dieser besten Wortfolge korrespondierenden unformatierten Textinformation TI' ausgebildet.

Die Erkennungsmittel 24 weisen weiters eine Formatierung-Speicherstufe 107 und eine Formatierung-Stufe 108 auf. Die Formatierung-Speicherstufe 107 ist zum Speichern einer Formatierung-Information FI ausgebildet, mit deren Hilfe Regeln repräsentierbar sind, die angeben, wie die unformatierte Textinformation TI' zu formatieren ist. Die Formatierung-Stufe 108 ist zum Empfangen der unformatierten Textinformation TI' und zum Zugreifen auf die Formatierung-Speicherstufe 107 und zum Auslesen der Formatierung-Information FI ausgebildet. Die Formatierung-Stufe 108 ist weiters unter Ausnutzung der Formatierung-Information FI zum Formatieren der unformatierten Textinformation TI' und als ein Ergebnis des Formatierens zum Erzeugen und zum Abgeben der Textinformation TI ausgebildet.

Die Erkennungsmittel 24 weisen weiters eine siebente Modell-Speichestufe 109 auf, die für jeden Empfangskanal und jede Sprache und jede Sprechergruppe zum Speichern von jeweils einem dritten Phonem-Modell PM3 ausgebildet und vorgesehen sind. Weiters ist eine siebente Modell-Auswahlstufe 110 vorgesehen, die zum Empfangen der Kanalangabe-Information CHI und der Sprach-Information LI und der Sprechergruppe-Information SGI ausgebildet ist. Die siebente Modell-Auswahlstufe 110 ist weiters unter Ausnutzung der Informationen CHI, LI und SGI zum Zugreifen auf das zu diesen Informationen CHI, LI und SGI korrespondierende dritte Phonem-Modell PM3 in der siebenten Modell-Speicherstufe 109 und zum Abgeben dieses kanal-, sprache- und sprechergruppespezifischen dritten Phonem-Modells PM3 an die Sprachmuster-Erkennungsstufe 105 ausgebildet. Die Erkennungsmittel 24 weisen weiters eine Kontext-Speicherstufe 111 auf. Die Kontext-Speicherstufe 111 ist zum Speichern der Kontext-Daten CD vorgesehen, welche Kontext-Daten CD zu jeder Kontext-Information CI und zu jeder Sprache eine Lexikon-Information LXI und eine zu der Lexikon-Information LXI korrespondierendes Sprache-Modell LM repräsentieren. Die Kontext-Speicherstufe 111 weist einen Lexikon-Speicherbereich 113 auf, in dem die jeweilige Lexikon-Information

LXI speicherbar ist, welche Lexikon-Information LXI Wörter und Phonem-Transkriptionen der Wörter umfasst. Die Kontext-Speicherstufe 111 weist einen Sprache-Modell-Speicherbereich 112 auf, in dem ein zu der jeweiligen Lexikon-Information LXI korrespondierendes Sprache-Modell LM speicherbar ist. Die Erkennungsmittel 24 weisen weiters eine Kontext-Auswahlstufe 114 auf, die zum Empfangen der Kontext-Information CI ausgebildet ist.

Es sei an dieser Stelle erwähnt, dass die Sprache-Information LI nicht explizit zu der Kontext-Auswahlstufe 114 zugeführt wird, weil die Kontext-Information CI die Sprache implizit repräsentiert.

Die Kontext-Auswahlstufe 114 ist weiters unter Ausnutzung der Kontext-Information CI und der damit implizit repräsentierten Information über die jeweilige Sprache zum Zugreifen auf das in der Kontext-Speicherstufe 111 zu der jeweiligen Kontext-Information CI korrespondierende Sprach-Modell LM bzw. auf die Lexikon-Information LXI und zum Abgeben des ausgewählten Sprache-Modells LM und der ausgewählten Lexikon-Information LXI in Form der Kontext-Daten CD an die Sprachmuster-Erkennungsstufe 105 ausgebildet.

Die Sprache-Erkennungsmittel 24 sind weiters auf trainierbare Weise zum Erzeugen der dritten Phonem-Modelle PM3 und der Lexikon-Information LXI und dem jeweils zu einer Lexikon-Information LXI korrespondierenden Sprach-Modell LM ausgebildet. Die siebente Modell-Speicherstufe 109 und die Kontext-Speicherstufe 111 bilden in diesem Zusammenhang trainierbare Ressourcen der Erkennungsmittel 24.

Zum Zweck des Trainierens der trainierbaren Ressourcen weisen die Erkennungsmittel 24 eine siebente Datenstromsteuerstufe 115 und eine siebente Trainingsstufe 116 auf. Die siebente Datenstromsteuerstufe 115 ist im Fall des Trainings dazu ausgebildet, die Sprache repräsentierenden Featurevektoren FV nicht an die Sprachmuster-Erkennungsstufe 105, sondern an die siebente Trainingsstufe 116 abzugeben. Die siebente Trainingsstufe 116 ist zum Empfangen der Sprache repräsentierenden Featurevektoren FV und der dazu korrespondierenden Training-Text-Information TTI ausgebildet. Die siebente Trainingsstufe 116 ist weiters unter Ausnutzung der Featurevektoren FV und der Training-Text-Informationen TTI und unter Zuhilfenahme eines Viterbi-Algorithmus zum Erzeugen und zum Abgeben des jeweiligen dritten Phonem-Modells PM3 an die siebente Modell-Auswahlstufe 110 ausgebildet, so dass das

dritte trainierte Phonem-Modell PM3, das zu dem der Kanalangabe-Information CHI bzw. zu der Sprache-Information LI bzw. zu der Sprachegruppe-Information SGI korrespondiert mit Hilfe der siebente Modell-Auswahlstufe 110 in der siebenten Modell-Speicherstufe 109 an einer durch die Informationen CHI, SGI und LI definierten Speicherposition speicherbar ist.

Die Ermittlungsmittel 24 weisen weiters eine Sprache-Modell-Trainingsstufe 117 auf, die zum Empfangen eines durch eine Corpora-Information COR repräsentierten relativ großen Trainingstextes, der im Fachjargon als Corpora bezeichnet wird, ausgebildet ist. Die Sprache-Modell-Trainingsstufe 117 ist unter Ausnutzung der Corpora-Information COR und unter Zuhilfenahme des durch die Information CI angegebenen Topic und der implizit durch die Information CI angegebenen Sprache bestimmten Lexikon-Information LXI zum Trainieren bzw. zum Erzeugen des zu jeder Kontext-Information CI und der damit implizit repräsentierten Sprache korrespondierenden Sprache-Modells LM ausgebildet, wobei die dermaßen bestimmte Lexikon-Information LX1 mit Hilfe der Kontext-Auswahlstufe 114 aus Lexikon-Speicherstufe 113 auslesbar und an die Sprache-Modell-Trainingsstufe 117 abgebbar ist. Die Sprache-Modell-Trainingsstufe 117 ist zum Abgeben der trainierten Sprachmodelle LM an die Kontext-Auswahlstufe 114 ausgebildet, wonach das Sprach-Modell LM mit Hilfe der Kontext-Auswahlstufe 114 unter Ausnutzung der Information CI an dem jeweils dafür vorgesehenen Speicherplatz der Sprache-Modell-Speicherbereichs 112 gespeichert wird.

Die Erkennungsmittel 24 weisen weiters eine Lexikon-Erzeugungsstufe 118 auf, die ebenfalls zum Empfangen der Corpora-Information COR und unter Ausnutzung der Corpora-Information COR zum Erzeugen und zum Abgeben einer zu jeder Kontext-Information CI und der damit implizit repräsentierten Sprache korrespondierenden Lexikon-Information LXI an die Kontext-Auswahlstufe 114 ausgebildet ist, wonach die Lexikon-Information LXI mit Hilfe der Kontext-Auswahlstufe 114 unter Ausnutzung der Information CI an dem jeweils dafür vorgesehenen Speicherplatz der Lexikon-Speicherbereich 112 gespeichert wird. Zum Zweck des Erzeugens der Lexikon-Information LXI weisen die Erkennungsmittel 24 eine Hintergrundlexikon-Speicherstufe 119 auf, die zum Speichern eines Hintergrundlexikons ausgebildet ist, welches Hintergrundlexikon einen Grundstock von Wörtern und dazu gehörenden phonetischen Transkriptionen von Wörtern aufweist, die repräsentiert durch eine Hintergrund–Transkription-Information BTI

abgebbar ist. Die Erkennungsmittel 24 wiesen weiters eine Statistik-Transkription-Stufe 120 auf, die auf Grundlage eines statistischen Transkriptionsverfahrens zum Erzeugen einer phonetischen Transkription von in dem Trainingstext enthaltenen Wörtern ausgebildet ist, die repräsentiert durch eine Statistik-Transkription-Information STI abgebbar ist.

Die Erkennungsmittel 24 weisen weiters eine Phonetik-Transkriptionsstufe 121 auf, die zum Empfangen jedes einzelnen Wortes des Trainingstexts enthaltenden Corpora-Text-Information CTI und unter Berücksichtigung der Kontext-Information CI und der implizit enthaltenen Information über die Sprache und zum Bereitstellen bzw. zum Abgeben einer phonetischen Transkription jedes Wortes der Corpora-Text-Information CTI in Form einer Corpora-Phonetik-Transkription-Information CPTI für die Lexikon-Erzeugungsstufe 118 ausgebildet ist. Zu diesem Zweck ist die Phonetik-Transkriptionsstufe 121 zum Prüfen ausgebildet, ob in der Hintergrundlexikon-Speicherstufe 119 eine geeignete phonetische Transkription für das jeweilige Wort verfügbar ist. Trifft dies zu, so bildet die Information BTI die Information CPTI. Ist keine geeignete Transkription verfügbar, so ist die Phonetik-Transkriptionsstufe 121 zum Bereitstellen der das jeweilige Wort repräsentierenden Information STI als die Information CTI ausgebildet.

An dieser Stelle sei erwähnt, dass die dritten Phonem-Modelle PM3 auch als akustische Referenzen bezeichnet werden, so dass die trainierbaren Ressourcen die akustischen Referenzen und den Kontext umfassen.

Es sei an dieser Stelle erwähnt, dass bei der Stufe 69, 80, 97 und 116 jeweils ein sogenanntes Training-Lexikon zum Einsatz kommt, mit dessen Hilfe aus der Training-Text-Information TTI eine für das jeweilige Training notwendige phonetische Transkription erzeugt wird.

Die auf mehrstufige Weise erzeugbaren und jeweils eine Spracheigenschaft repräsentierenden Informationen ASI, LI, SGI und CI bewirken bei den Sprach-Erkennungsmitteln 24 im wesentlichen drei Effekte. Gemäß einem ersten Effekt wird bei der dritten Sprachfilterstufe 104 mit Hilfe der Segmentierung-Information ASI das Filtern der Featurevektoren FV gesteuert. Dadurch ist der Vorteil erhalten, dass das Erkennen der Textinformation TI autonom und unabhängig von einer vorhergehende Beeinflussung – beispielsweise durch ein Hintergrundgeräusch - der die Sprachinformation SI repräsentierenden Featurevektoren FV auf präzise und rasche Weise durchführbar ist.

Gemäß einem zweiten Effekt wird bei den Ressourcen mit Hilfe der Kanalangabeinformation CHI und der Sprache-Information LI und der Sprechergruppe-Information SGI das Auswählen einer zu diesen Informationen korrespondierenden akustischen Referenz gesteuert. Dadurch ist der Vorteil erhalten, dass ein wesentlicher Beitrag zu dem präzisen Erkennen der Textinformation TI erhalten ist, weil die akustische Referenz die akustischen Spracheigenschaften der Sprache mit hoher Genauigkeit modelliert. Gemäß einem dritten Effekt wird bei den Ressourcen mit Hilfe der Kontext-Information das Auswählen eines Kontexts gesteuert. Dadurch ist der Vorteil erhalten, dass ein weiterer positiver Beitrag zu einem präzisen und raschen Erkennen der Textinformation TI erhalten ist. Hinsichtlich des präzisen Erkennens ist der Vorteil deshalb erhalten, weil ein auswählbarer Kontext den tatsächlich bei einer Sprache vorliegenden Kontext viel genauer modelliert als dies im Falle eines starr vorgegebenen relativ großen Kontexts der Fall wäre. Hinsichtlich des raschen Erkennens ist der Vorteil deshalb erhalten, weil der jeweilige zu einer der Kontext-Information CI korrespondierende Wortschatz nur einen Teil der Wörter einer Sprache abdeckt und daher relativ klein sein kann und daher entsprechend rasch verarbeitbar ist.

Im vorliegenden Fall hat es sich als vorteilhaft erwiesen, dass die Erkennungsstufen 21, 22 und 24 jeweils eine eigene Sprachfilterstufe 61, 76 und 104 aufweisen. Die Erkennungsstufe 23 enthält wegen ihrer Funktionalität implizit eine Sprachfilterung. Es sei erwähnt, dass an Stelle der drei Sprachfilterstufen 61, 76 und 104 auch eine einzige in der Figur 1 dargestellte Sprachfilterstufe 122 vorgesehen sein kann, die den Erkennungsstufen 21, 22, 23 und 24 vorgeschaltet ist, was aber die Funktionalität der Erkennungsstufe 23 nicht beeinträchtigt. Dadurch wäre der Vorteil erhalten, dass die drei Sprachfilterstufen 61, 76 und 104 nicht notwendig sind und daher auch die Verarbeitung der Featurevektoren FV unter Umständen beschleunigt durchführbar ist.

Es sei erwähnt, dass an Stelle des den Mittel 20 bis 24 vorgeschalteten Featurevektor-Extrahierungsmittels 19 jedes der Mittel 20 bis 24 ein ihm zugeordnetes individuelles Featurevektor-Extrahierungsmittel aufweisen kann, dam das vorverarbeitete Audiosignal PAS zuführbar ist. Dadurch ist ermöglicht, dass jedes der individuellen Featurevektor-Extrahierungsmittel an die Funktion des jeweiligen Mittel 20 bis 24 optimal individuell angepasst sein kann. Dadurch ist der Vorteil erhalten, dass die vektormäßige Repräsentation des vorverarbeiteten Audiosignals PAS individuelle angepasst auch auf

einer anderen als der Cepstral-Ebene stattfinden kann.

Es sei erwähnt, dass die Sprachinformation SI der Spracherkennungseinrichtung 1 auch mit Hilfe eines Speichermediums oder unter Zuhilfenahme eines Computernetzwerks verfügbar gemacht werden kann.

Es sei erwähnt, dass die Stufe 12 auch durch Hardware realisiert sein kann.

Es sei erwähnt, dass die Umwandlungsstufen-Erzeugungsstufe 16 auch durch eine Hardwarelösung realisiert sein kann.

Es sei erwähnt, dass die Teilbereiche des Audiosignals PAS und die dazu korrespondierenden Informationen CHI, ASI, LI, SGI und CI auch als sogenannte Softwareobjekte speicherbar sein können und dass die Erkennungsmittel 18, 20, 21, 22, 23 und 24 zum Erzeugen, Verändern und zum Verarbeiten dieser Softwareobjekte ausgebildet sein können. Weiters kann vorgesehen sein, dass das Speichern der Teilbereiche des Audiosignals PAS und das Speichern bzw. das Verwalten der jeweils zugehörigen Informationen CHI, ASI, LI, SGI und CI selbstständig von den Mitteln 18, 20, 21, 22, 23, 24 und 25 durchgeführt werden kann. Es sei weiters erwähnt, dass die Mittel 8, 19 und die Stufe 122 durch ein Softwareobjekt realisiert sein kann. Gleiches gilt auch für die Erkennungsmittel 18, 20, 21, 22, 23, 24 und 25. Weiters sei erwähnt, dass die Mittel 8, 18, 19, 20, 21, 22, 23, 24 und 25 und die Stufe 122 auch durch Hardware realisiert sein können.

Das Mittel 24 realisiert in dem vorstehend erläuterten Ausführungsbeispiel einen sogenannten „Large Vocabulary Continuous Speech Recogniser". Es sei jedoch erwähnt, dass die Mittel 24 auch einen sogenannten „Command and Control Recogniser" realisieren können, wobei in diesem Fall der Kontext nur aus einem Lexikon ohne ein Sprache-Modell besteht. Weiters sind zusätzlich Maßnahmen vorgesehen, die ein Verwalten von mindestens einem Grammatik-Modell erlauben.

Für die Zwecke der Mittel 23 und 24 kann auch vorgesehen sein, dass die Information CHI, LI und SGI zu einer sogenannten Phonem-Modell-Information zusammengefasst sind, weil die drei Informationen das jeweilige Phonem-Modell bestimmen, obwohl die Information LI bei dem Mittel 23 unabhängig und zusätzlich zu der Phonem-Modell-Information verwendet wird. Dadurch ist der Vorteil erhalten, dass die Architektur der Spracherkennungseinrichtung 1 vereinfacht ist.

Weiters kann vorgesehen sein, dass bei den Mittel 20 zusätzlich ein Erkennen

von sogenannten „Hesitations“ vorgesehen sein kann.

Patentansprüche:

1. Spracherkennungseinrichtung zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation, welche Sprachinformation hinsichtlich von Spracheigenschaften charakterisierbar ist,
wobei erste Spracheigenschaft-Erkennungsmittel vorgesehen sind, die unter Ausnutzung der Sprachinformation zum Erkennen einer ersten Spracheigenschaft und zum Erzeugen einer die erkannte erste Spracheigenschaft repräsentierenden ersten Eigenschaftsinformation ausgebildet sind, und
wobei zumindest zweite Spracheigenschaft-Erkennungsmittel vorgesehen sind, die unter Ausnutzung der Sprachinformation zum Erkennen einer zweiten Spracheigenschaft der Sprachinformation und zum Erzeugen einer die erkannte zweite Spracheigenschaft repräsentierenden zweiten Eigenschaftsinformation ausgebildet ist, und
wobei Sprach-Erkennungsmittel vorgesehen sind, die unter fortwährender Berücksichtigung von zumindest der ersten Eigenschaftsinformation und der zweiten Eigenschaftsinformation zum Erkennen der zu der Sprachinformation korrespondierenden Textinformation ausgebildet sind.

2. Spracherkennungseinrichtung nach Anspruch 1,
wobei Empfangsmittel vorgesehen sind, die zum Empfangen der Sprachinformation über mindestens zwei erkennbare Empfangskanäle ausgebildet sind, und
wobei Empfangskanal-Erkennungsmittel vorgesehen sind, die zum Erkennen des jeweils zum Empfangen der Sprachinformation verwendeten Empfangskanals und zum Erzeugen einer den erkannten Empfangskanal repräsentierenden Kanalangabe-Information ausgebildet sind, und
wobei mindestens eines der mindestens zwei Spracheigenschaft-Erkennungsmittel oder/und die Sprach-Erkennungsmittel zum Berücksichtigen der Kanalangabe-Information ausgebildet ist.

3. Spracherkennungseinrichtung nach Anspruch 1,
wobei die Sprach-Erkennungsmittel zeitverzögert um mindestens eine Zeitspanne, die von den mindestens zwei Spracheigenschaft-Erkennungsmitteln zum Erzeugen der mindestens zwei Eigenschaftsinformationen benötigt wird und während der ein Teil der Sprachinformation von den mindestens zwei Spracheigenschaft-Erkennungsmitteln zum Erzeugen der mindestens zwei Eigenschaftsinformationen ausgenutzt wird, zum Erkennen

der Textinformation ausgebildet sind, die zumindest zu einem Teilbereich des zum Erzeugen der mindestens zwei zugeführten Eigenschaftsinformationen ausgenutzten Teils der Sprachinformation korrespondieren.

4. Spracherkennungseinrichtung nach Anspruch 1,
wobei mindestens eine mit Hilfe von Spracheigenschaft-Erkennungsmitteln erzeugte Eigenschaftsinformation anderen Spracheigenschaft-Erkennungsmitteln zuführbar ist und wobei die anderen Spracheigenschaft-Erkennungsmittel beim Erkennen der Spracheigenschaft der Sprachinformation und beim Erzeugen der Eigenschaftsinformation zum Berücksichtigen der mindestens einen zugeführten Eigenschaftsinformation ausgebildet sind.

5. Spracherkennungseinrichtung nach Anspruch 4,
wobei die anderen Spracheigenschaft-Erkennungsmittel, zeitverzögert um mindestens eine Zeitspanne, die zum Erzeugen der mindestens einen zugeführten Eigenschaftsinformationen benötigt wird und während der ein Teil der Sprachinformation von den Spracheigenschaft-Erkennungsmitteln zum Erzeugen der mindestens einen zugeführten Eigenschaftsinformationen ausgenutzt wird, zum Erkennen der Spracheigenschaft ausgebildet sind, die zumindest einen Teilbereich des zum Erzeugen der mindestens einen zugeführten Eigenschaftsinformation ausgenützten Teils der Sprachinformation charakterisiert.

6. Spracherkennungsverfahren zum Erkennen einer zu einer Sprachinformation korrespondierenden Textinformation, welche Sprachinformation hinsichtlich von Spracheigenschaften charakterisierbar ist,
wobei unter Ausnutzung der Sprachinformation eine erste Spracheigenschaft erkannt wird und
wobei eine die erkannte erste Spracheigenschaft repräsentierende erste Eigenschaftsinformation erzeugt wird und
wobei unter Ausnutzung der Sprachinformation mindestens eine zweite Spracheigenschaft erkannt wird und
wobei eine die erkannte zweite Spracheigenschaft repräsentierende zweite Eigenschaftsinformation erzeugt wird und
wobei die zu der Sprachinformation korrespondierende Textinformation unter fortwährender Berücksichtigung von zumindest der ersten Eigenschaftsinformation und der

zweiten Eigenschaftsinformationen erkannt wird.

7. Spracherkennungsverfahren nach Anspruch 6,

wobei die Sprachinformation über einen Empfangskanal von mindestens zwei erkennbaren Empfangskanälen empfangen wird und

wobei der jeweils zum Empfangen der Sprachinformation verwendete Empfangskanal erkannt und eine den erkannten Empfangskanal repräsentierende Kanalangabe-Information erzeugt wird und

wobei zumindest bei dem Erzeugen von mindestens einer der Eigenschaftsinformationen oder/und bei dem Erkennen der Textinformation die Kanalangabe-Information berücksichtigt wird.

8. Spracherkennungsverfahren nach Anspruch 6,

wobei das Erkennen der zu der Sprachinformation korrespondierenden Textinformation zeitverzögert um mindestens eine Zeitspanne , die zum Erzeugen der mindestens zwei Eigenschaftsinformationen benötigt wird und während der ein Teil der Sprachinformation zum Erzeugen der mindestens zwei Eigenschaftsinformationen ausgenützt wird, für die zumindest zu einem Teilbereich des zum Erzeugen der mindestens zwei Eigenschaftsinformationen ausgenutzten Teils der Sprachinformation korrespondierende Textinformation erfolgt.

9. Spracherkennungsverfahren nach Anspruch 6,

wobei mindestens eine Spracheigenschaft unter Berücksichtigung von mindestens einer nicht diese Spracheigenschaft repräsentierenden Eigenschaftsinformation erkannt wird und eine die erkannte Spracheigenschaft repräsentierende Eigenschaftsinformation erzeugt wird.

10. Spracherkennungsverfahren nach Anspruch 9, dadurch gekennzeichnet,

wobei das Erkennen der mindestens einen Spracheigenschaft unter Berücksichtigung von mindestens einer nicht diese Spracheigenschaft repräsentierenden Eigenschaftsinformation zeitverzögert um mindestens eine Zeitspanne, die zum Erzeugen der mindestens einen nicht diese Spracheigenschaft repräsentierenden Eigenschaftsinformation benötigt wird und während der ein Teil der Sprachinformation zum Erzeugen der mindestens einen nicht diese Spracheigenschaft repräsentierenden Eigenschaftsinformation ausnutzbar ist, für zumindest einen Teilbereich des zum Erzeugen der mindestens einen nicht diese Spracheigenschaft repräsentierenden Eigenschaftsinformation ausgenützten Teils der

Sprachinformation erfolgt.

11. Computerprogrammprodukt,

das direkt in einen Speicher eines Computers geladen werden kann und Softwarecodeabschnitte umfasst, wobei mit dem Computer das Spracherkennungsverfahren gemäß dem Anspruch 6 abgearbeitet werden kann, wenn das Computerprogrammprodukt auf dem Computer abgearbeitet wird.

12. Computerprogrammprodukt nach Anspruch 11,

wobei das Computerprogrammprodukt auf einem computerlesbaren Medium gespeichert ist.

13. Computer mit einer Recheneinheit und einem internen Speicher, welcher Computer das Computerprogrammprodukt gemäß dem Anspruch 11 abarbeitet.

Zusammenfassung

Spracherkennungseinrichtung mit Mitteln
zum Berücksichtigen von mindestens zwei Spracheigenschaften

Bei einer zum Erkennen einer zu einer Sprachinformation (SI) korrespondierenden Textinformation (TI), wobei die Sprachinformation (SI) hinsichtlich von Spracheigenschaften charakterisierbar ist, sind erstens mindestens zwei Spracheigenschaft-Erkennungsmittel (20, 21, 22, 23) vorgesehen, wobei jedes der Spracheigenschaft-Erkennungsmittel (20, 21, 22, 23) unter Ausnutzung der Sprachinformation (SI) zum Erkennen einer ihm zugeordneten Spracheigenschaft und zum Erzeugen einer die erkannte Spracheigenschaft repräsentierenden Eigenschaftsinformation (ASI, LI, SGI, CI) ausgebildet ist, und sind zweitens Sprach-Erkennungsmittel (24) vorgesehen, die unter fortwährender Berücksichtigung der mindestens zwei Eigenschaftsinformationen (ASI, LI, SGI, CI) zum Erkennen der zu der Sprachinformation (SI) korrespondierenden Textinformation (TI) ausgebildet sind.

(Figur 1).

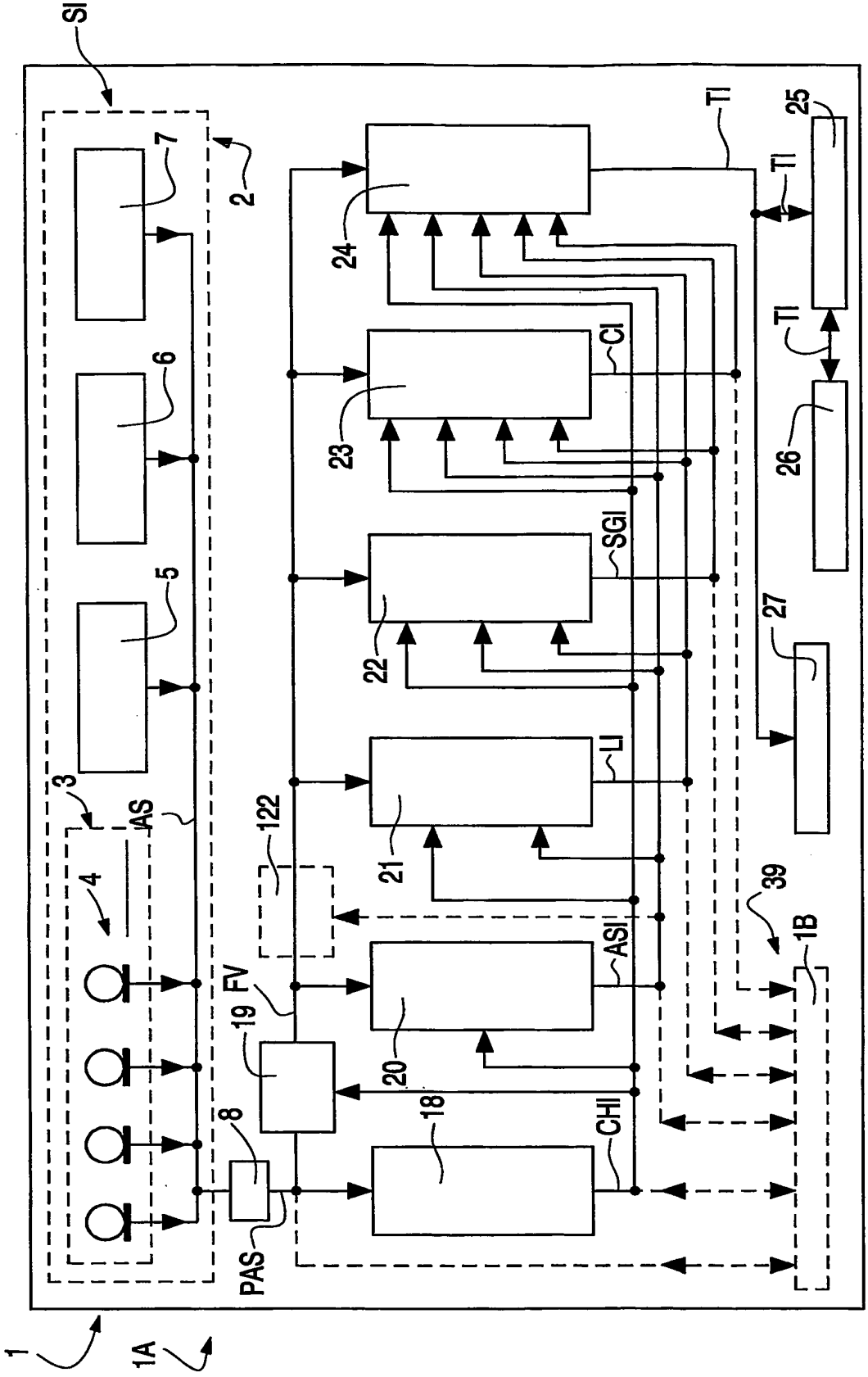
SI
2
7
6
5
3
AS
4
122
FV
19
8
PAS
1
1A
24
23
22
21
20
18
TI
CI
SGI
LI
ASI
CHI
25
26
27
39
1B

Fig.1

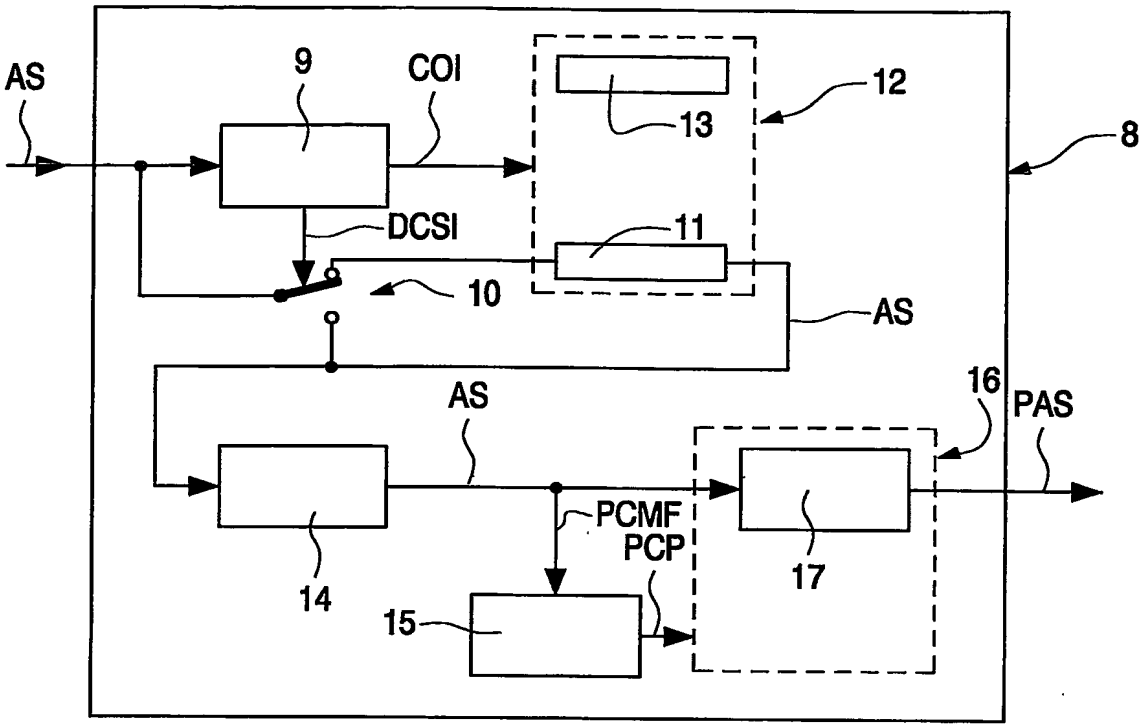
AS
9
COI
12
13
8
DCSI
11
10
AS
AS
16
PAS
PCMF
PCP
14
15
17

Fig.2

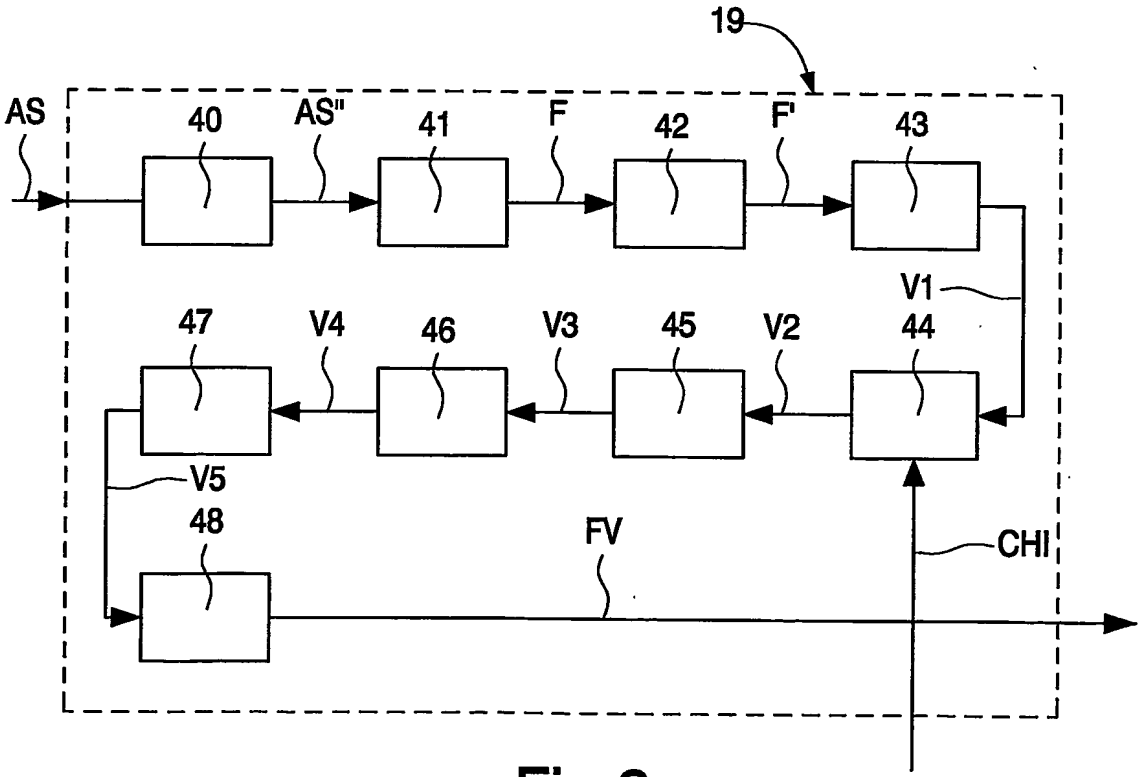
19
AS
40
AS''
41
F
42
F'
43
V1
47
V4
46
V3
45
V2
44
V5
48
FV
CHI

Fig.3

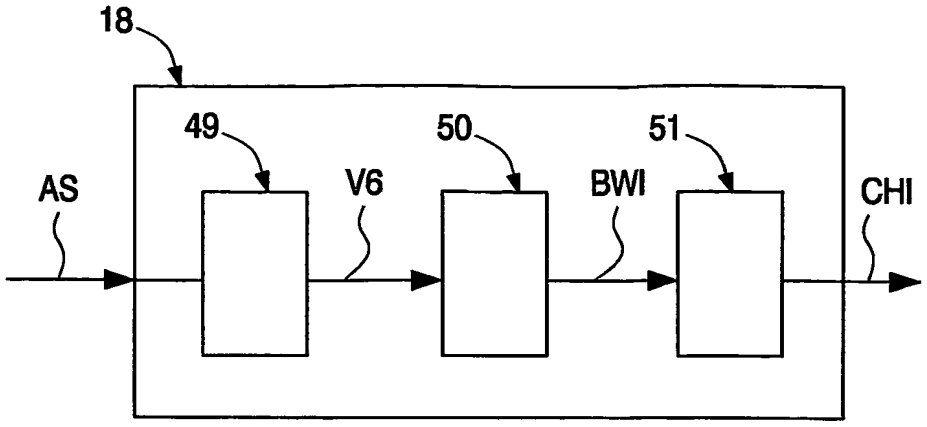

Fig.4

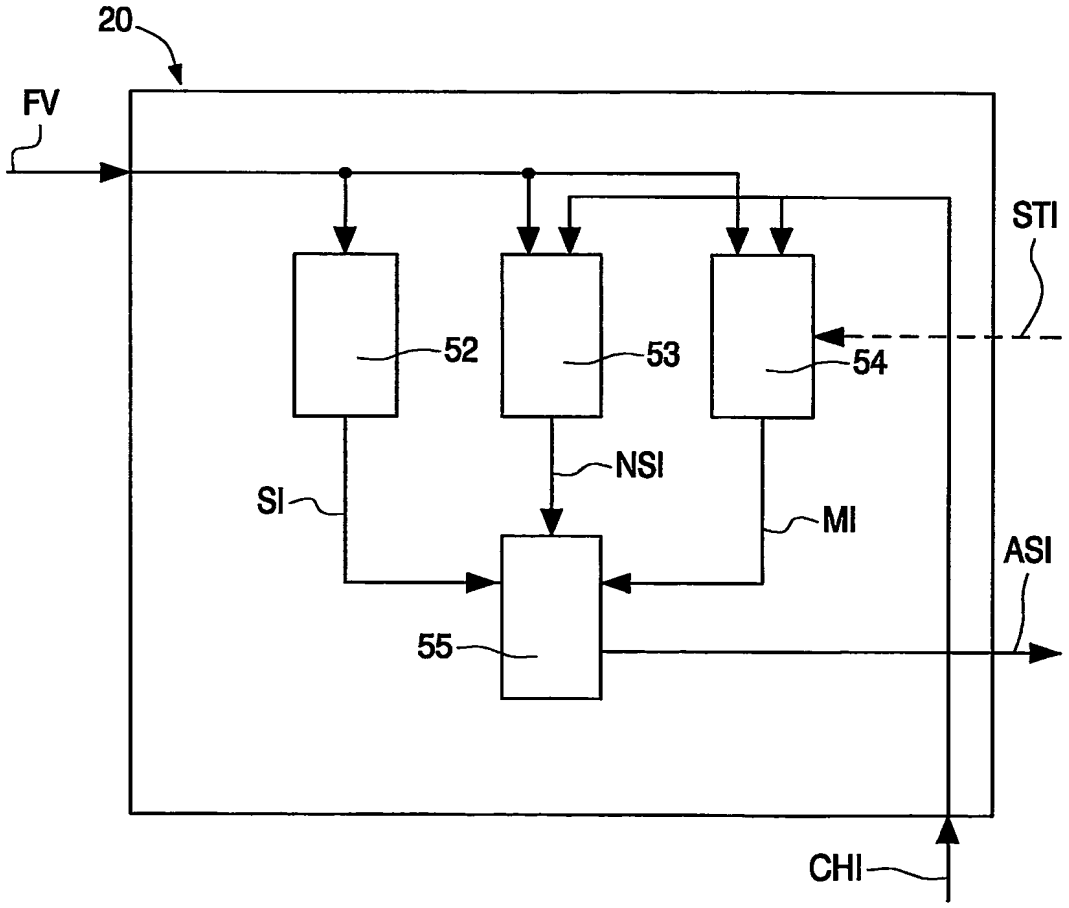

Fig.5

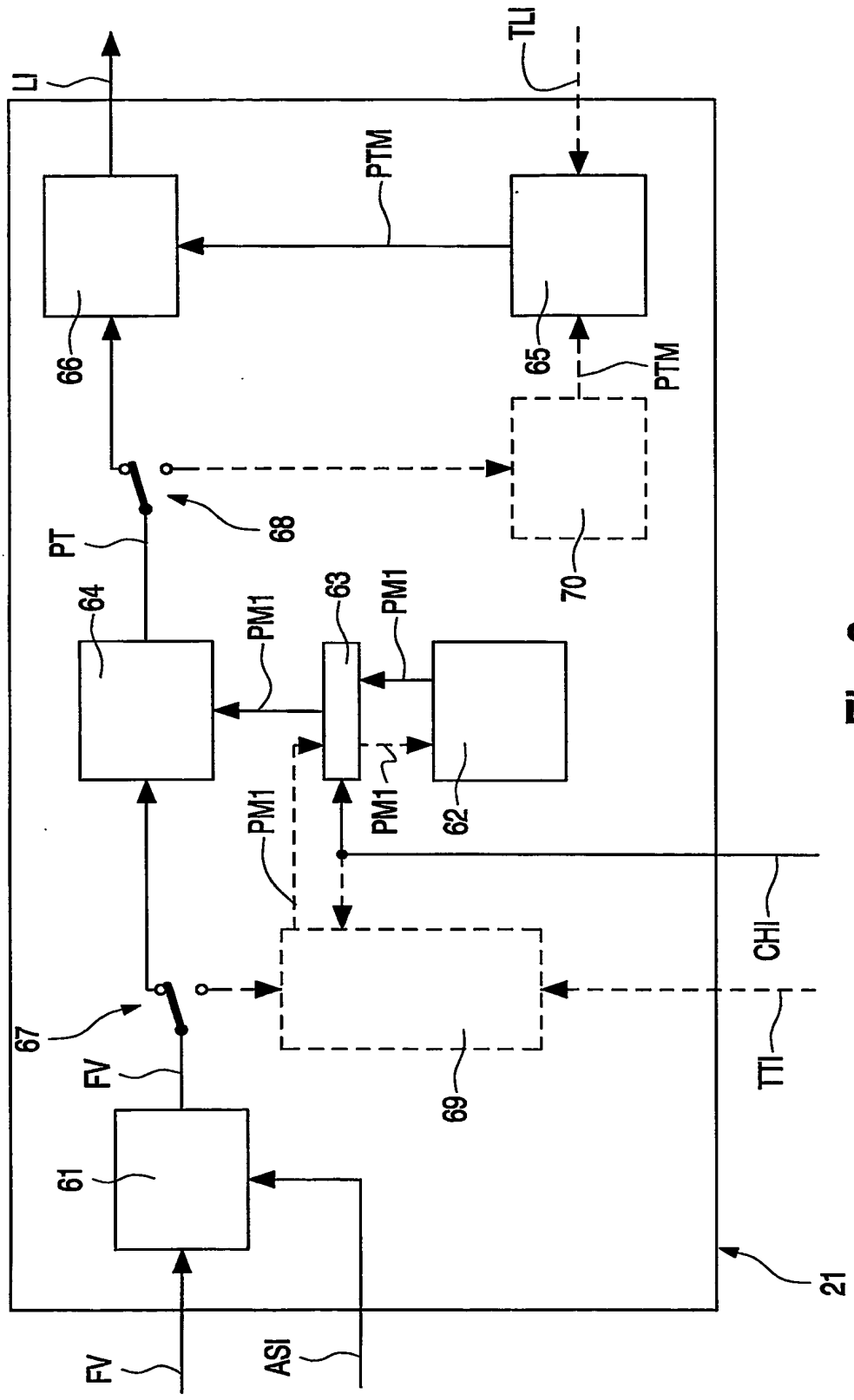
LI
TLI
PTM
66
65
PTM
PT
68
70
64
PM1
63
PM1
62
PM1
PM1
CHI
67
FV
69
TTI
61
21
FV
ASI

Fig.6

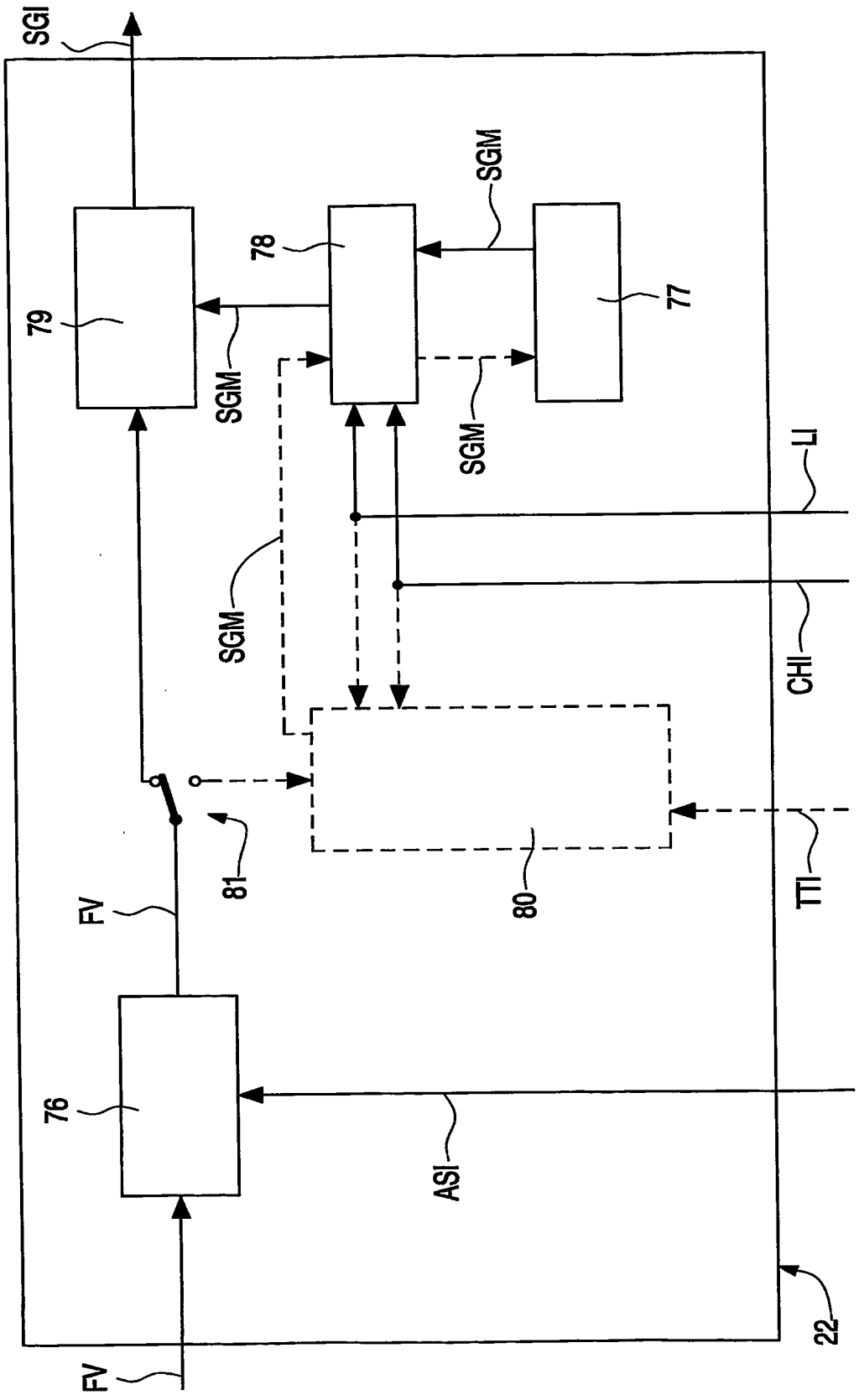

Fig.7

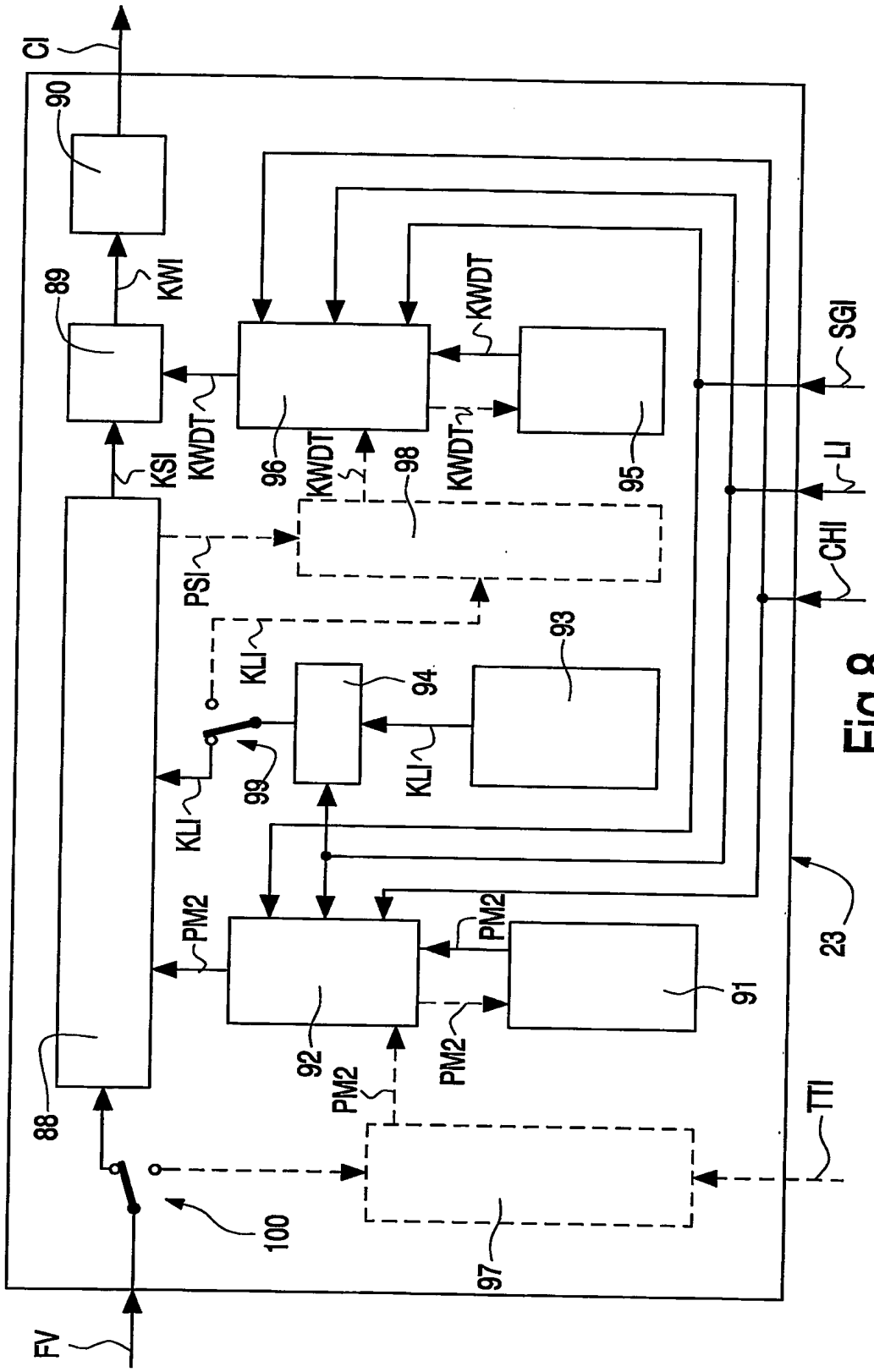
CI
90
89
KWI
KSI
KWDT
96
KWDT
98
KWDT
95
KWDT
SGI
LI
CHI
PSI
KLI
93
94
KLI
99
KLI
PM2
PM2
PM2
PM2
PM2
88
92
91
23
TTI
100
97
FV

Fig.8

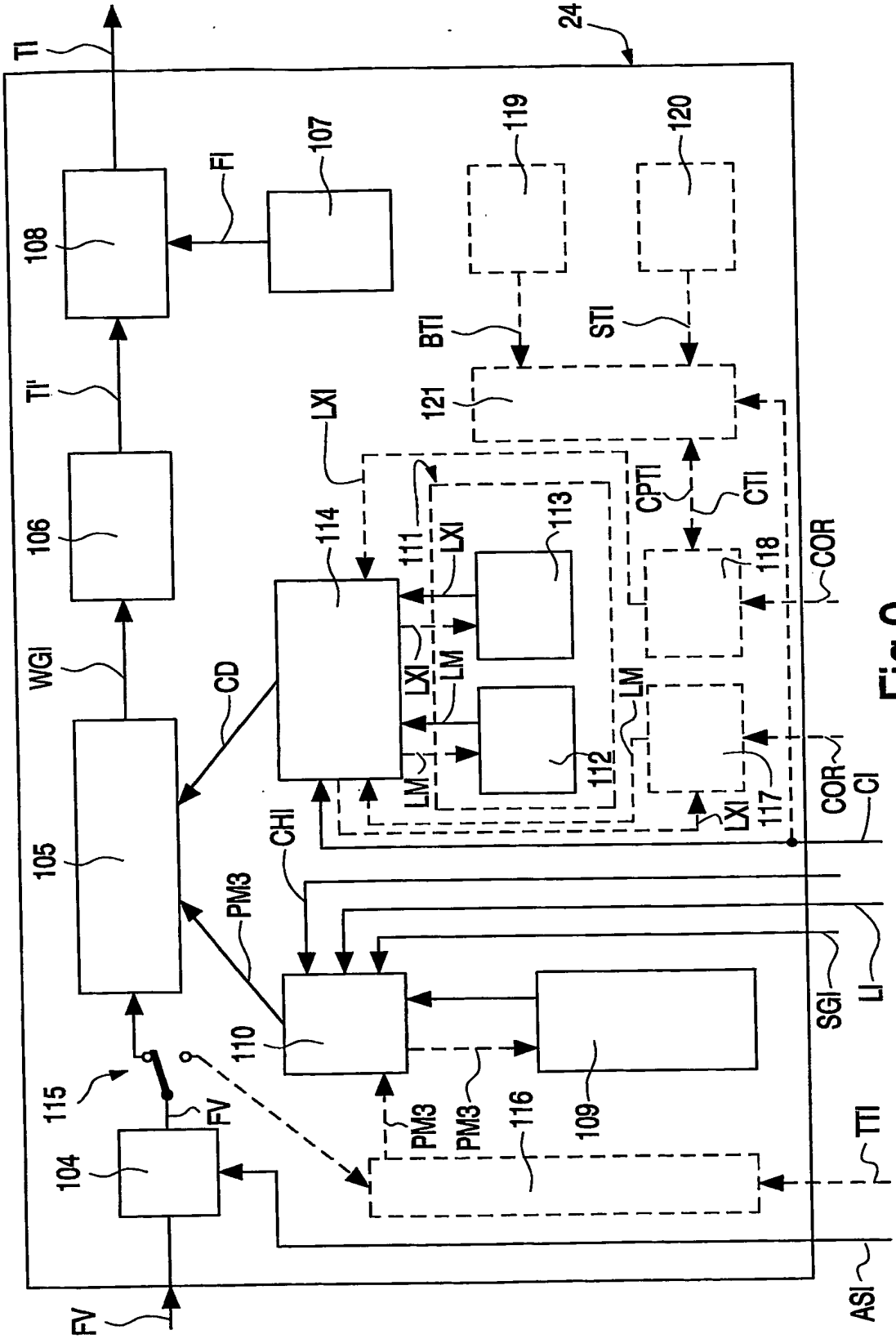
24
TI
FI
107
108
119
120
BTI
STI
TI'
121
LXI
106
114
111
113
CPTI
CTI
118
COR
WGI
CD
LM
112
117
CI
105
CHI
PM3
110
SGI
LI
115
FV
104
116
109
TTI
ASI
FV

Fig.9

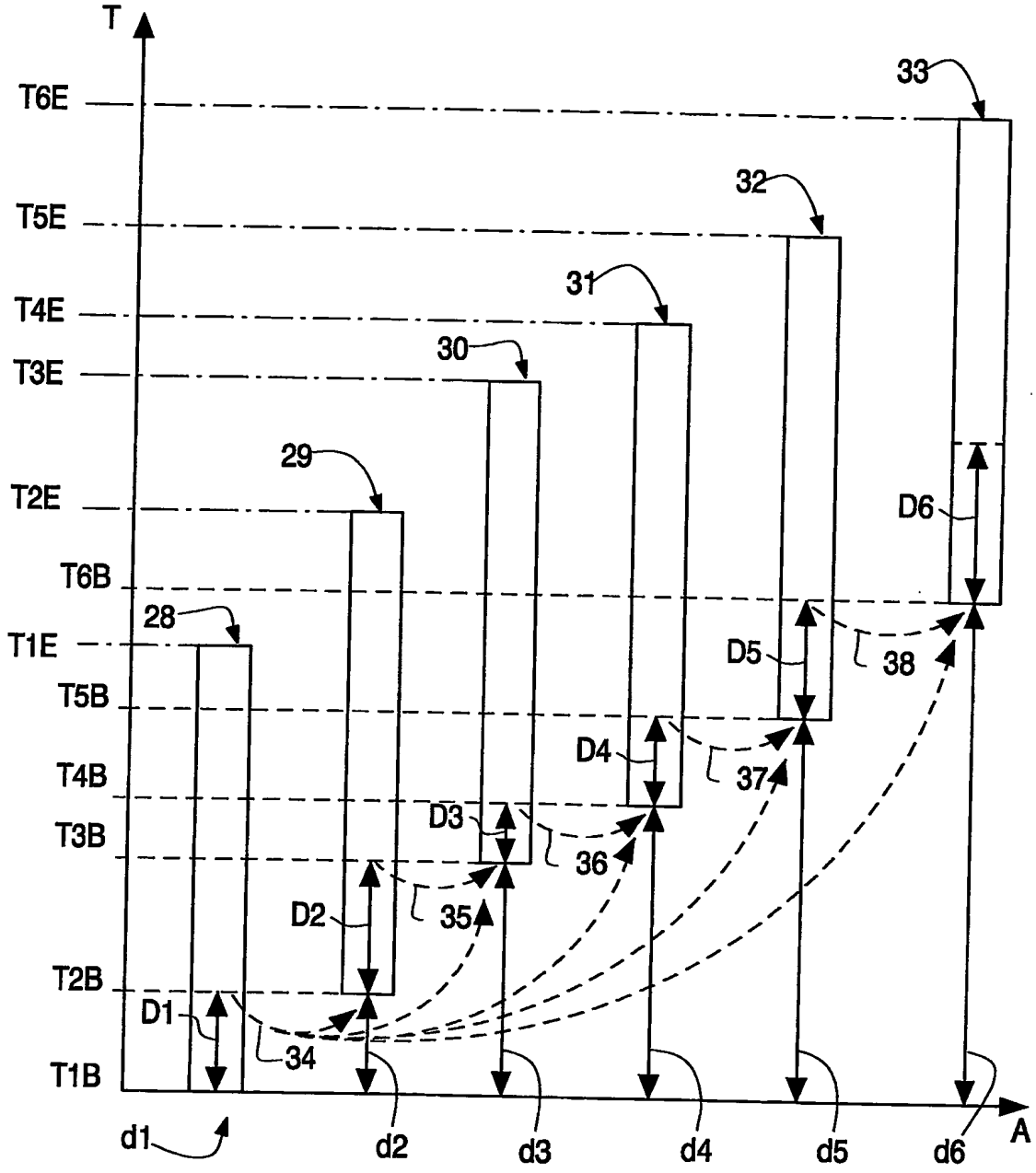
T
T6E
T5E
T4E
T3E
T2E
T6B
T1E
T5B
T4B
T3B
T2B
T1B
A
d1
d2
d3
d4
d5
d6
28
29
30
31
32
33
D1
D2
D3
D4
D5
D6
34
35
36
37
38

Fig.10

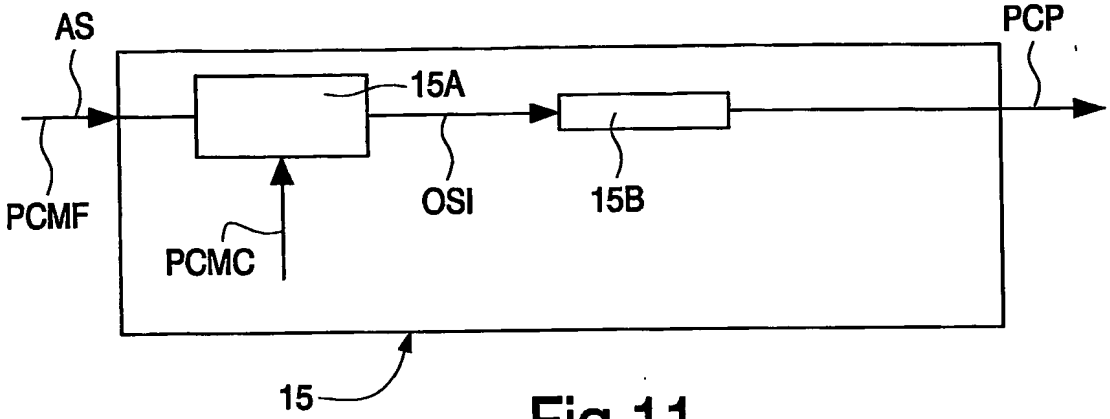

Fig.11

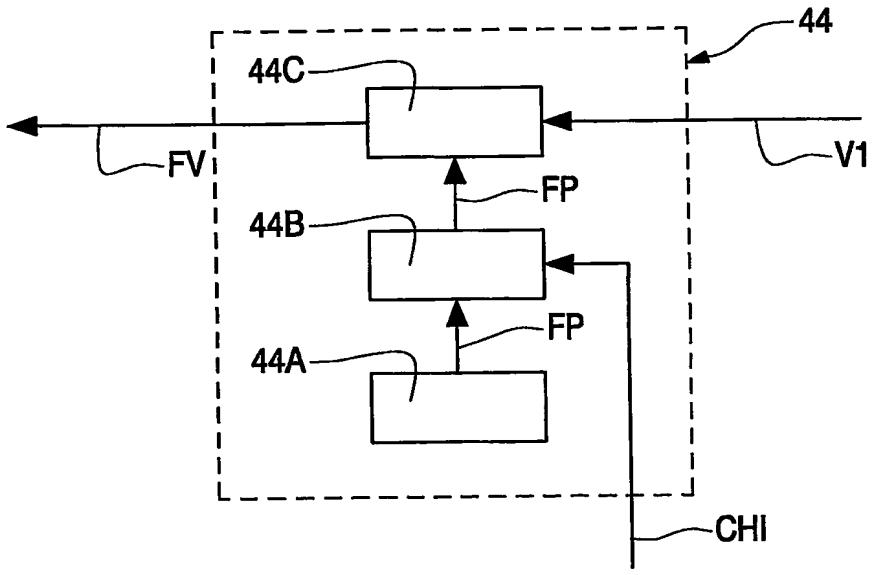

Fig.12

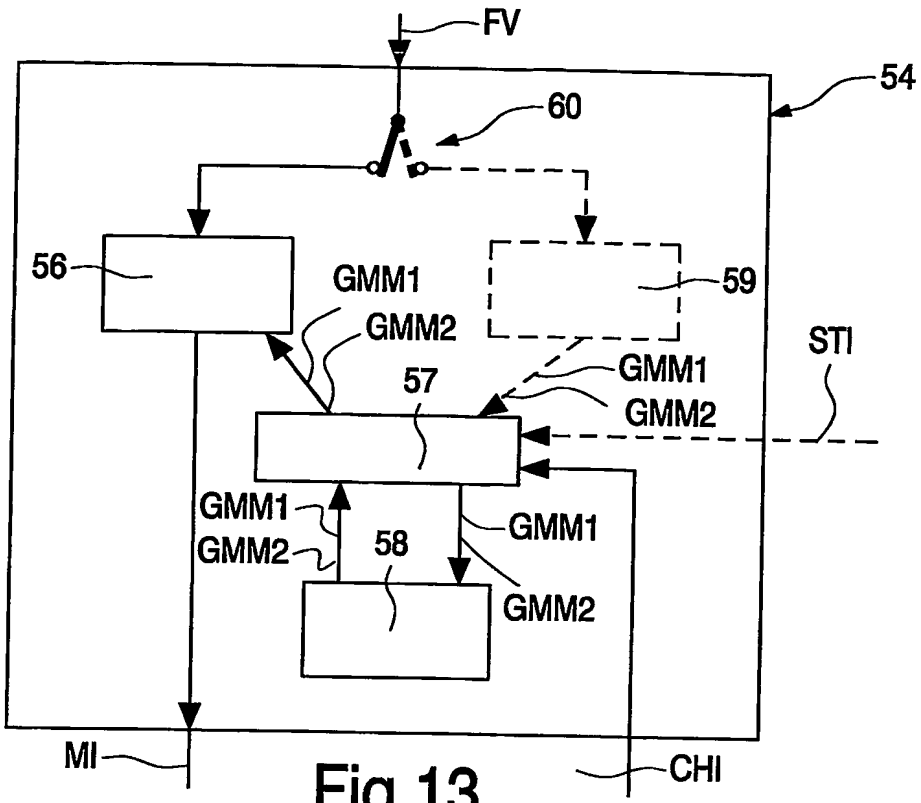
FV
54
60
56
GMM1
GMM2
59
STI
57
GMM1
GMM2
GMM1
GMM2
58
GMM1
GMM2
MI
CHI

Fig.13

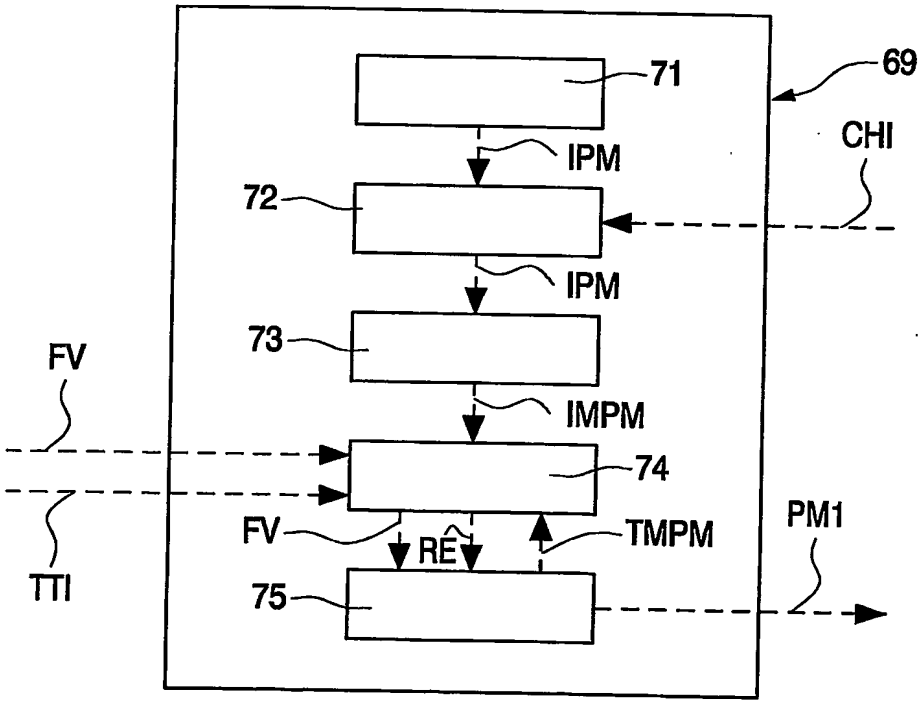
71
69
IPM
CHI
72
IPM
73
FV
IMPM
74
FV
RE
TMPM
PM1
TTI
75

Fig.14

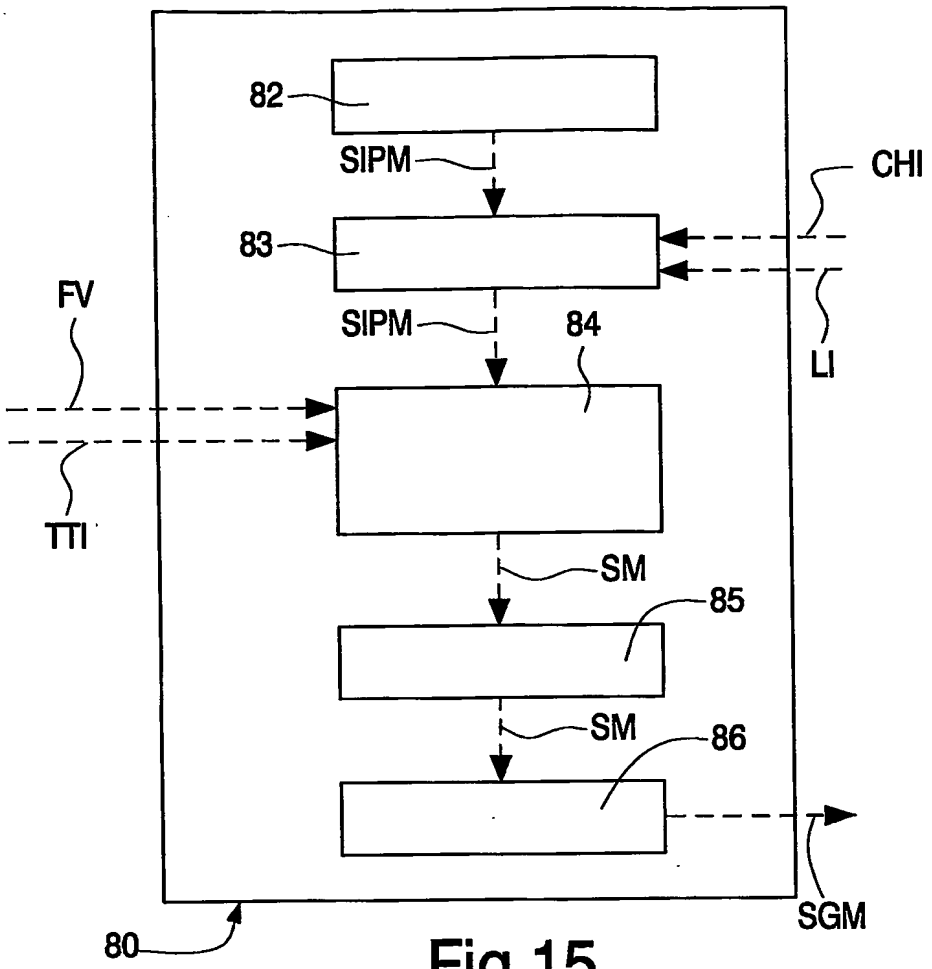
82
SIPM
CHI
83
FV
SIPM
84
LI
TTI
SM
85
SM
86
SGM
80

Fig.15

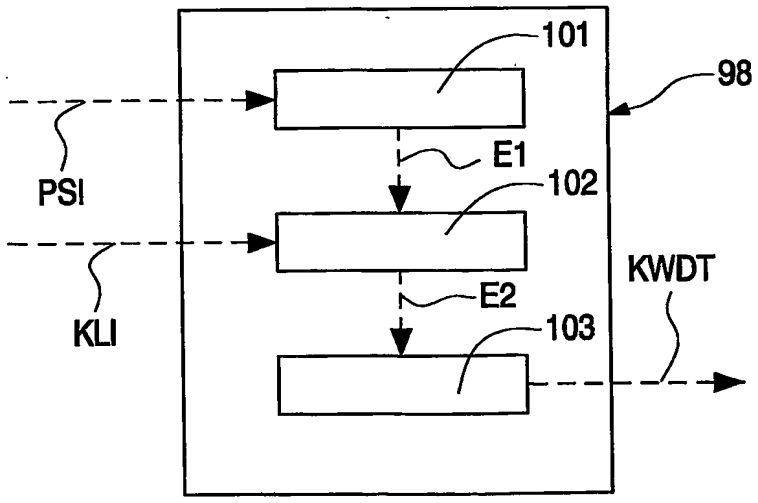
101
98
PSI
E1
102
KLI
KWDT
E2
103

Fig.16